

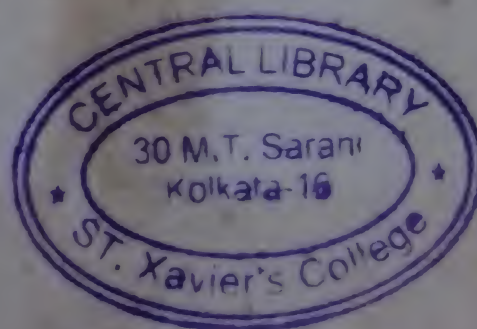






85.2
K
9 (3)

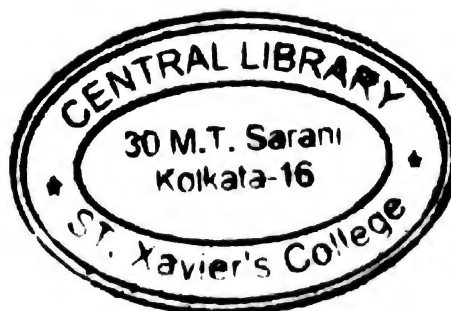
DEPARTMENT OF STATISTICS
St Xavier's College
Calcutta.



THE ADVANCED THEORY
OF STATISTICS

VOLUME 3

*DESIGN AND ANALYSIS,
AND TIME-SERIES*



OTHER BOOKS ON STATISTICS AND MATHEMATICS

- | | |
|---|------------------------------|
| <i>Combinatorial chance</i> | F. N. DAVID and D. E. BARTON |
| <i>Games, gods and gambling</i> | F. N. DAVID |
| <i>A statistical primer</i> | F. N. DAVID |
| <i>An introduction to the theory of statistics</i> | G. U. YULE and M. G. KENDALL |
| <i>Rank correlation methods</i> | M. G. KENDALL |
| <i>Exercises in theoretical statistics</i> | M. G. KENDALL |
| <i>Exercises in probability and statistics (for mathematics undergraduates)</i> | N. A. RAHMAN |
| <i>Rapid statistical calculations</i> | M. H. QUENOUILLE |
| <i>Biomathematics</i> | C. A. B. SMITH |
| <i>The design and analysis of experiment</i> | M. H. QUENOUILLE |
| <i>Sampling methods for censuses and surveys</i> | F. YATES |
| <i>Random processes and the growth of firms</i> | J. STEINDL |
| <i>Statistical method in biological assay</i> | D. J. FINNEY |
| <i>The mathematical theory of epidemics</i> | N. T. J. BAILEY |
| <i>Probability and the weighing of evidence</i> | I. J. GOOD |

GRIFFIN'S STATISTICAL MONOGRAPHS AND COURSES:

- | | |
|---|----------------------------------|
| No. 1: <i>The analysis of multiple time-series</i> | M. H. QUENOUILLE |
| No. 2: <i>A course in multivariate analysis</i> | M. G. KENDALL |
| No. 3: <i>The fundamentals of statistical reasoning</i> | M. H. QUENOUILLE |
| No. 4: <i>Basic ideas of scientific sampling</i> | A. STUART |
| No. 5: <i>Characteristic functions</i> | E. LUKACS |
| No. 6: <i>An introduction to infinitely many variates</i> | E. A. ROBINSON |
| No. 7: <i>Mathematical methods in the theory of queueing</i> | A. Y. KHINTCHINE |
| No. 8: <i>A course in the geometry of n dimensions</i> | M. G. KENDALL |
| No. 9: <i>Random wavelets and cybernetic systems</i> | E. A. ROBINSON |
| No. 10: <i>Geometrical probability</i> | M. G. KENDALL and P. A. P. MORAN |
| No. 11: <i>An introduction to symbolic programming</i> | P. WEGNER |
| No. 12: <i>The method of paired comparisons</i> | H. A. DAVID |
| No. 13: <i>Statistical assessment of the life characteristic:
a bibliographic guide</i> | W. R. BUCKLAND |
| No. 14: <i>Applications of characteristic functions</i> | E. LUKACS and R. G. LAHA |
| No. 15: <i>Elements of linear programming with economic applications</i> | R. C. GEARY and M. D. MCCARTHY |
| No. 16: <i>Inequalities on distribution functions</i> | H. J. GODWIN |
| No. 17: <i>Green's function methods in probability theory</i> | J. KEILSON |
| No. 18: <i>The analysis of variance</i> | A. HUITSON |
| No. 19: <i>The linear hypothesis: a general theory</i> | G. A. F. SEBER |
| No. 20: <i>Econometric techniques and problems</i> | C. E. V. LESER |
| No. 21: <i>Stochastically dependent equations</i> | P. R. FISK |

Descriptive brochure available from Charles Griffin & Co. Ltd.

THE ADVANCED THEORY OF STATISTICS

MAURICE G. KENDALL, M.A., Sc.D.

*Managing Director, C-E-I-R Ltd. • Formerly Professor of Statistics in the
University of London • President of the Royal Statistical Society, 1960-2*

and

ALAN STUART, D.Sc. (ECON.)

Professor of Statistics in the University of London

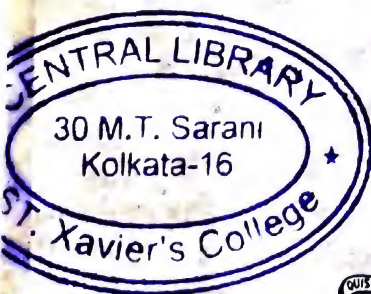
CENTRAL LIBRARY
St. Xavier's College, Kolkata

IN THREE VOLUMES

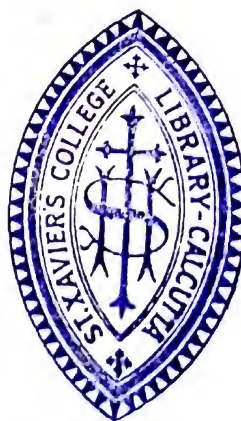
DEPARTMENT OF STATISTICS
St Xavier's College
Calcutta.

VOLUME 3

DESIGN AND ANALYSIS,
AND TIME-SERIES



A-3572



CHARLES GRIFFIN & COMPANY LIMITED
LONDON



Copyright © 1966
CHARLES GRIFFIN & COMPANY LIMITED,
42, DRURY LANE, LONDON, W.C.2

85.2

K
9 (3)

THE ADVANCED THEORY OF STATISTICS

Volume 1		Volume 2	
First published ..	1943	First published ..	1946
Second edition ..	1945	Second edition ..	1947
Third ..	1947	Third ..	1951
Fourth ..	1948	Second impression	1955
Fifth ..	1952	Third ..	1959
		Fourth ..	1960

Three-volume edition

Volume 1: Distribution Theory

First published .. 1958
Second edition .. 1963

Volume 2: Inference and Relationship

First published .. 1961

Volume 3: Design and Analysis, and Time-Series

First published .. 1966

71 392

28 MAR 1967

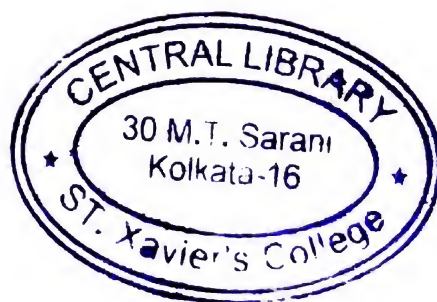
Made and printed in Great Britain by
Butler & Tanner Ltd., Frome and London

DEPARTMENT OF STATISTICS
St Xavier's College
Calcutta.

ACC. NO. A-3572
CALL NO. 519.5 KEN(ADV).V3
BOOK NO. 71392
DATE 28/03/1967

"... the race is not to the swift, nor the battle to the strong, neither yet bread to the wise, nor yet riches to men of understanding, nor yet favour to men of skill; but time and chance happeneth to them all."

Ecclesiastes, 9. 11



h
it
y
th
I
in
h
M

it
to
L
e

fo

en
bo

LO
A

PREFACE TO VOLUME 3

This is the final volume of our treatise. It has taken longer to write than we had hoped. To some extent this has been due to our involvement with other work, but it is also attributable to the amount of development which has been going on in recent years in the subjects dealt with in this volume, which are the analysis of variance, the design of experiments, sample survey theory, multivariate analysis, and time-series. It becomes increasingly difficult to know what is permanent and what is ephemeral in the spate of current research. In deciding what to omit and what to admit, there have been occasions when we have been reminded of what Thackeray said about Macaulay, that he read a book to write a sentence.

As with the first two volumes, this one is self-contained in three respects: it lists its own references, it contains such Appendix Tables as are necessary to follow the text, and it has its own index. Now that the Kendall-Doig *Bibliography of Statistical Literature* is available, a comprehensive bibliography is unnecessary. As before, extensive sets of exercises are provided at the ends of chapters.

We have again to thank Mr. E. V. Burke of Charles Griffin and Company Limited for the care he has given to the production of this work.

We are also grateful to many reviewers and correspondents who have commented on errors, misprints and obscurities in the first two volumes, and shall be equally glad to be notified of any that may be found in this final volume.

M. G. K.
A. S.

LONDON

August, 1966

0
3
3
3
3
3
3
4
4
12
13
4
5
6
7
3
1
1

CONTENTS

<i>Chapter</i>	<i>Page</i>
	1
35. Analysis of Variance in the Linear Model: Classified Data	57
36. Other Models for the Analysis of Variance	85
37. The Assumptions of the Analysis of Variance	119
38. The Design of Experiments	166
39. Sample Survey Theory: Designs	211
40. Sample Survey Theory: Supplementary Information	239
41. Multivariate Distribution Theory	264
42. Tests of Hypotheses in Multivariate Analysis	285
43. Canonical Variables	314
44. Discrimination and Classification	342
45. Time-Series: General	366
46. Time-Series: Trend and Seasonality	403
47. Stationary Time-Series	431
48. The Sampling Theory of Serial Correlations	454
49. Spectrum Theory	472
50. Time-Series: Some Further Topics	504
Envoi	505
Appendix Tables	517
References	539
Index	

CHAPTER 35

ANALYSIS OF VARIANCE IN THE LINEAR MODEL: CLASSIFIED DATA

35.1 In developing the MV unbiased linear estimation properties of the LS estimator (19.12, Vol. 2) in the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ at (19.8), we observed at (19.42) that the sum of squares (SS) of the observations may be written identically as the sum of two non-negative components

$$\mathbf{y}'\mathbf{y} \equiv (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}) + (\mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{X}\hat{\boldsymbol{\theta}}) \quad (35.1)$$

of which the first is the sum of squared residuals (Residual SS) from the model fitted by LS. The second component on the right of (35.1) is the reduction in the SS due to the fitted model; the greater this reduction is (i.e. the smaller the Residual SS is), the more satisfactorily the fitted model represents the \mathbf{y} - \mathbf{X} relationship in the observations. If we rewrite (35.1) as

$$\mathbf{y}'\mathbf{y} \equiv (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}) + \hat{\boldsymbol{\theta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\theta}} \quad (35.2)$$

and recall from (19.16) that

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad (35.3)$$

we see that if the error-vector $\boldsymbol{\epsilon}$ in the model is normally distributed, $\hat{\boldsymbol{\theta}}$, being a linear function of $\boldsymbol{\epsilon}$, will by 15.4 be multinormally distributed with mean $\boldsymbol{\theta}$ and dispersion matrix (35.3). The last term on the right of (35.2) is therefore the exponent of this multinormal distribution apart from the factor $-(2\sigma^2)^{-1}$, and by 24.6, $\hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}}/\sigma^2$ is distributed in the non-central χ^2 form (24.18) with degrees of freedom $\nu = k$ and non-central parameter

$$\lambda = \boldsymbol{\theta}'\mathbf{V}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{\theta} = \boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta}/\sigma^2. \quad (35.4)$$

For brevity we write this distribution $\chi'^2(\nu, \lambda)$ as in 24.5.

35.2 This result enables us to test the hypothesis that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and in particular to test

$$H_0 : \boldsymbol{\theta} = \mathbf{0}, \quad (35.5)$$

for then λ defined by (35.4) is zero, and the distribution becomes a (central) χ^2 with k degrees of freedom (d.fr.). As we saw in 19.11-12, $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})/\sigma^2$ is a χ^2 with $(n - k)$ d.fr., and $\mathbf{y}'\mathbf{y}/\sigma^2$ is a χ^2 with n d.fr. when (35.5) holds, Cochran's theorem of 15.16 applies, and the two components on the right of (35.2) are independently distributed. Their ratio (after division by d.fr.)

$$F = \{\hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}}/k\} / \{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})/(n - k)\} \quad (35.6)$$

has the variance-ratio F -distribution with $(k, n - k)$ d.fr. when (35.5) holds.

If we wish to investigate the power of a test of (35.5) based on (35.6), we require its distribution when (35.5) does not hold. In order to show that it is a non-central F as

at (24.105), we must prove that the numerator and denominator of (35.6) remain independent when $\theta \neq 0$. If we wish to test the more general hypothesis that $\theta = \theta_0 \neq 0$, we require this distribution in order to make a test at all. Thus we need an extension of Cochran's theorem (15.16) to non-central normal variables, i.e. normal variables with means not all equal.

35.3 Apart from this particular need, the form of (35.2) is suggestive in another way. Suppose that $\mathbf{X}'\mathbf{X}$ is a diagonal matrix, say \mathbf{C} , with diagonal elements c_{ii} . The last term on the right of (35.2) can then be further separated into

$$(\mathbf{X}\hat{\theta})'(\mathbf{X}\hat{\theta}) = \sum_{i=1}^k c_{ii} \hat{\theta}_i^2. \quad (35.7)$$

The elements c_{ii} are positive, since $\mathbf{X}'\mathbf{X}$ is a positive definite (non-singular) matrix. (35.7) expresses the reduction in the SS as the sum of k parts, one corresponding to each parameter. Here again, we may be interested in testing hypotheses concerning individual θ_i , and require the distribution of the components $c_{ii} \hat{\theta}_i^2$ when $\theta \neq 0$.

If $\mathbf{X}'\mathbf{X}$ is diagonal, so is (35.3), and the linear model is called *orthogonal* since the θ_i are uncorrelated, and actually independent when ϵ is normal. We have already discussed orthogonal models in the context of regression theory in 28.15–20, Vol. 2, where we were concerned with the use of orthogonal polynomials. (28.72) defined (and Example 28.3 illustrated) the procedure of evaluating the reduction in the SS due to each further parameter, using an entirely intuitive justification. Our present discussion will be more general.

Analysis of variance

35.4 We now introduce a fundamental concept, originally developed by R. A. Fisher in the 1920's. If the SS $\mathbf{y}'\mathbf{y}$ can be expressed as the sum of non-negative components, each of which corresponds to a subset of the parameters of the linear model, we call this an *analysis of variance* (AV) on y . (It would be more appropriate to call it an analysis of SS, but history and brevity are against this logical usage.) An AV is interpreted as a separating-out of the influences of the different subsets of parameters upon the observations \mathbf{y} . The importance of such separations in many fields of investigation make AV the central technique of much applied statistics.

Decomposition of non-central quadratic forms

35.5 We now state a general AV problem. Suppose that \mathbf{y} is a vector of p independent normal variates with

$$E(\mathbf{y}) = \boldsymbol{\mu}, \quad V(\mathbf{y}) = \mathbf{I}, \quad (35.8)$$

and that

$$\mathbf{y}'\mathbf{y} = \sum_{i=1}^k \mathbf{y}'\mathbf{A}_i\mathbf{y}, \quad (35.9)$$

where \mathbf{A}_i has rank r_i . Under what conditions will the k quadratic forms $Q_i = \mathbf{y}'\mathbf{A}_i\mathbf{y}$ be independently distributed, and what will their distributions be? Since, by 24.4, the distribution of $\mathbf{y}'\mathbf{y} = Q$ is a $\chi^2(p, \boldsymbol{\mu}'\boldsymbol{\mu})$, we may expect the components to have the same distributional form.

This differs from the problem considered in 15.16 only in that there we had $\boldsymbol{\mu} = 0$.

We saw there that any one of the three conditions

- (a) $\sum_{i=1}^k r_i = \text{the rank of } Q$,
- (b) each A_i is idempotent, i.e. $A_i = A_i^2$,
- (c) $A_i A_j = 0$, all $i \neq j$,

implies the other two. Re-examination of the proof of this in 15.17-19 will reveal that it did not depend on the value of μ at all. Neither did the proof (in 15.13) of Craig's theorem that Q_i and Q_j are independent if and only if $A_i A_j = 0$, which shows that (c) is equivalent to

- (c') each Q_i is independent of every other.

However, the equivalence of (b) and the statement

- (b°) each Q_i is a (central) χ^2 variable with r_i d.fr.,

depended upon the result of 15.11 that if $\mu = 0$, Q_i is a χ^2 variable if and only if A_i is idempotent. It is thus (b°) which requires to be generalized through a generalization of 15.11 to $\mu \neq 0$.

35.6 The only essential change brought about in 15.11 is that the canonically transformed variable y_i^2 in (15.43) is now a $\chi'^2(1, \mu_i^2)$ variable, by 24.4. The c.g.f. of $\sum_{i=1}^r a_i y_i^2$ is therefore not (15.45), but the more general form obtained from Exercise 24.1, which yields for the cumulants of Q

$$\kappa_s = 2^{s-1}(s-1)! \sum_{i=1}^r a_i^s (1 + s\mu_i^2) \quad (35.10)$$

(the generalization of (15.46)), and also shows that the cumulants of a $\chi'^2(\nu, \lambda)$ variable are

$$\kappa_s^* = 2^{s-1}(s-1)! (\nu + s\lambda). \quad (35.11)$$

For (35.10) and (35.11) to be identical, we must have

$$\left. \begin{aligned} \sum_{i=1}^r a_i^s &= \nu, \\ \sum_{i=1}^r a_i^s \mu_i^2 &= \lambda, \end{aligned} \right\} \text{all } s. \quad (35.12)$$

(35.12) is satisfied if and only if every $a_i = 1$, so that $\nu = r$ and $\lambda = \sum_{i=1}^r \mu_i^2$. Since the a_i are the non-zero latent roots of A , it follows that A is idempotent. We thus see that, for general μ , the statement equivalent to (b) above is

- (b') each Q_i is a $\chi'^2(r_i, \lambda_i)$ variable,

reducing to (b°) when $\mu = 0$. Moreover, if we transform orthogonally back from the canonical to the original variables, we see at once that $\lambda_i = \mu' A_i \mu$, and $\sum_{i=1}^k \lambda_i = \mu' \mu$, the non-central parameters of the Q_i adding to that of Q (cf. Exercise 24.1).

We have thus reached the conclusion that if (35.8-9) hold, any of the conditions (a), (b), (c) implies the other two; equivalently, any of the conditions (a), (b') and (c') implies the other two of them.

35.7 The result of 35.6 is unaffected if, in (35.9), $\mathbf{y}'\mathbf{y}$ is replaced by $Q = \mathbf{y}'\mathbf{A}\mathbf{y}$, where \mathbf{A} is any idempotent matrix with rank $r < p$. The argument justifying this in Chapter 15 for the case $\boldsymbol{\mu} = \mathbf{0}$ is valid for general $\boldsymbol{\mu}$.

Even if $\mathbf{V}(\mathbf{y})$ in (35.8) is generalized so that the components of \mathbf{y} have a non-singular multinormal distribution with dispersion matrix \mathbf{V} which is non-diagonal, the result is only slightly changed. For if

$$E(\mathbf{y}) = \boldsymbol{\mu}, \quad \mathbf{V}(\mathbf{y}) = \mathbf{V}, \quad (35.13)$$

and

$$Q = \mathbf{y}'\mathbf{A}\mathbf{y} = \sum_{i=1}^k \mathbf{y}'\mathbf{A}_i\mathbf{y} = \sum_{i=1}^k Q_i, \quad (35.14)$$

we may write $\mathbf{V} = \mathbf{T}\mathbf{T}'$ and the transformation $\mathbf{y} = \mathbf{T}\mathbf{z}$ produces independent normal variables \mathbf{z} , since the exponent of the multinormal distribution is

$$\mathbf{y}'\mathbf{V}^{-1}\mathbf{y} = \mathbf{z}'\mathbf{T}'(\mathbf{T}')^{-1}\mathbf{T}^{-1}\mathbf{T}\mathbf{z} = \mathbf{z}'\mathbf{z}.$$

We then have, from (35.14),

$$Q = \mathbf{z}'\mathbf{T}'\mathbf{A}\mathbf{T}\mathbf{z} = \sum_{i=1}^k \mathbf{z}'\mathbf{T}'\mathbf{A}_i\mathbf{T}\mathbf{z} = \sum_{i=1}^k Q_i,$$

and these are the quadratic forms with which we now deal. Condition (b) of 35.5 is now

$$\mathbf{T}'\mathbf{A}_i\mathbf{T} = \mathbf{T}'\mathbf{A}_i\mathbf{T}.\mathbf{T}'\mathbf{A}_i\mathbf{T}$$

or

$$\mathbf{A}_i\mathbf{V} = \mathbf{A}_i\mathbf{V}\mathbf{A}_i\mathbf{V}$$

so that $\mathbf{A}_i\mathbf{V}$ must now be idempotent, as must $\mathbf{A}\mathbf{V}$. Condition (c) is

$$\mathbf{T}'\mathbf{A}_i\mathbf{T}.\mathbf{T}'\mathbf{A}_j\mathbf{T} = \mathbf{0}$$

or

$$\mathbf{A}_i\mathbf{V}\mathbf{A}_j = \mathbf{0}, \quad i \neq j.$$

Condition (a) is unaffected by orthogonal transformation.

We may therefore finally state the general result:

If \mathbf{y} is non-singularly multinormal with moments (35.13), and the decomposition (35.14) holds for a quadratic form Q where $\mathbf{A}\mathbf{V}$ is idempotent, then Q is a $\chi'^2(r, \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu})$ variable, where r is the rank of \mathbf{A} , and any one of the three following conditions implies the others:

- (a) $\sum_{i=1}^k r_i = r$.
- (b) Each $\mathbf{A}_i\mathbf{V}$ is idempotent; equivalently, each Q_i is $\chi'^2(r_i, \boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu})$, where r_i is the rank of \mathbf{A}_i .
- (c) $\mathbf{A}_i\mathbf{V}\mathbf{A}_j = \mathbf{0}$; equivalently, each Q_i is independent of every other.

Graybill and Marsaglia (1957) give some more general results than this. Banerjee (1964) simplifies their proofs.

35.8 These general results on the decomposition of quadratic forms in normal variables solve the problems in 35.2-3, which motivated our investigation. For example, it now follows that the numerator of (35.6) is independent of the denominator, so that their ratio is duly distributed in the non-central F form [which we write, as in 24.31, $F'(\nu_1, \nu_2, \lambda)$] where $\nu_1 = k$, $\nu_2 = n - k$ and λ is given by (35.4) in this case.

Similarly, we now see that for any individual $\hat{\theta}_i$ in the orthogonal model of 35.3, the ratio

$$F = c_{ii}\hat{\theta}_i^2 / \{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})/(n-k)\} \quad (35.15)$$

is a $F'(1, n-k, c_{ii}\theta_i^2/\sigma^2)$ variable, and may then be used to test hypotheses concerning θ_i .

More generally, for a hypothesis H_0 imposing $r \leq k$ constraints, the ratio of the SS due to H_0 and the Residual SS, multiplied by $(n-k)/r$, is a $F'(r, n-k, \lambda)$ variable—cf. 24.29–31. It should be particularly noted from 24.29 that the non-central parameter λ is always of exactly the same form as the numerator SS of the test statistic with each observation replaced by its expectation, and σ^2 as a divisor. Thus we may always obtain λ very simply from the numerator SS in the test statistic by substituting $\boldsymbol{\theta}$ for $\hat{\boldsymbol{\theta}}$ and dividing by σ^2 .

These are examples of the LR test in the linear model, derived generally through the canonical form of the model in 24.25–9. The discussion below (24.100) explicitly pointed out that the LR test of a linear hypothesis concerning any subset of r of the k parameters is based upon the reduction in the SS due to these r parameters divided by the Residual SS. The canonical approach of Chapter 24 had its theoretical uses in the derivation of optimum properties of LR tests in 24.36–7. For our present purposes, the equivalent partitioning of SS which we have been discussing is more direct and informative.

We remind the reader that exact and approximate expressions for the power function of the LR F -tests are given in 24.32–3.

AV for classified observations

35.9 Our definition of AV in 35.4 applies to any linear model, and covers the applications to regression theory in 28.12–23. However, the term AV is commonly used in a narrower sense, in which it was originally developed.

We saw in 35.4 that AV is used to separate out the influences of different parameters upon y . In experimental work, the parameters are often the effects of certain “treatments” upon the variable y . For example, in agricultural experimentation, from which this terminology derives, y might be the yield of wheat from a plot of fixed size, and the “treatment” being investigated might be the addition of a certain fertilizer to the plot during the growing season. Naturally, the experiment would include both treated and untreated plots. The point here is that such an experiment may be brought within the scope of the general linear model by defining a “label” variable x which is equal to 1 when the treatment is given and 0 otherwise.

It is easy to see that any pattern of treatments can be handled in this way; we need only define a label variable x for each possible ingredient of the treatments in the experiment. If there are two fertilizers in the example of the previous paragraph, we should define x_1 as the label variable for the first and x_2 as the label variable for the second fertilizer. Thus, a plot which receives both fertilizers has $x_1 = x_2 = 1$; one which receives only the first has $x_1 = 1, x_2 = 0$; a plot which receives only the second fertilizer has $x_1 = 0, x_2 = 1$; and a plot receiving no fertilizer has $x_1 = x_2 = 0$. The analysis of the linear model can now proceed without difficulty, since the elements of \mathbf{X} may be any known constants.

35.10 The feature of the matrix \mathbf{X} in the examples discussed in 35.9 is that all its elements are units or zeros, since they are merely labels for the presence or absence of certain ingredients in the "treatments." In the narrower sense, the term AV is used to describe the analysis of a linear model when this restriction holds true for all the elements of \mathbf{X} . Other small positive integers are also permitted in \mathbf{X} in this narrower sense of AV. For example, in the single-fertilizer experiment discussed at the beginning of 35.9, some plots might be given a single dose, others a double dose, and others none at all of the fertilizer. We could then define $x = 2, 1$ or 0 accordingly; the analysis of this model could still be called AV. However, this formulation suffers from the fact that it implies that $E(y)$ is affected twice as much by a double as by a single dose—the model is linear in β , the "effect" parameter expressing the dependence of y upon x . This could be overcome by defining two label variables, x_1 to denote presence or absence of a single dose, and x_2 to denote presence or absence of a double dose of the fertilizer. This alternative formulation does not (as the reader may be tempted to think) reduce the model to the two-fertilizer model at the end of 35.9, for we cannot now have $x_1 = x_2 = 1$ for any plot—there is evidently some loss of symmetry to offset the avoidance of the implication of linearity in dose-effect.

35.11 We shall be discussing the formulation of linear models in several important AV situations, and we shall see that the simple (usually 0-1) structure of the elements of \mathbf{X} produces corresponding simplifications in the analysis itself. The simplest case is that of a classification of observations into groups, suspected to differ in their means; this is usually known as a *one-way classification*.

Example 35.1 AV in a one-way classification

Suppose that a sample of independent observations is classified into k groups, with n_i ($i = 1, 2, \dots, k$) observations in the i th group and $\sum_{i=1}^k n_i = n$. If the groups can differ only in their means, we may express this as

$y_{iq} = \theta_i + \varepsilon_{iq}$, $i = 1, 2, \dots, k$; $q = 1, 2, \dots, n_i$,
which is in the form of the general linear model

where

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

$$\mathbf{y}_{(n \times 1)} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{k1} \\ \vdots \\ y_{kn_k} \end{pmatrix}, \quad \boldsymbol{\theta}_{(k \times 1)} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix},$$

and

$$\mathbf{X}_{(n \times k)} = \begin{pmatrix} 1 & & & \\ \vdots & & & \\ 1 & & & \\ & 1 & & \\ & \vdots & & \\ & 1 & & \\ & & 1 & \\ & & \vdots & \\ & & 1 & \\ & & & \ddots \\ & & & 1 \\ & & & \vdots \\ & & & 1 \end{pmatrix} \begin{matrix} \left. \begin{matrix} 1 \\ \vdots \\ 1 \end{matrix} \right\} n_1 \text{ rows} \\ \left. \begin{matrix} 1 \\ \vdots \\ 1 \end{matrix} \right\} n_2 \text{ rows} \\ \left. \begin{matrix} 1 \\ \vdots \\ 1 \end{matrix} \right\} n_3 \text{ rows} \\ \vdots \\ \left. \begin{matrix} 1 \\ \vdots \\ 1 \end{matrix} \right\} n_k \text{ rows} \end{matrix}$$

The zero elements of \mathbf{X} are omitted. We see at once that

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n_1 & & & 0 \\ & n_2 & & \\ & & \ddots & \\ 0 & & & n_k \end{pmatrix},$$

so that the analysis is orthogonal (cf. 35.3) whatever the values of the n_i —in particular, they need not be equal. Also

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum_q y_{1q} \\ \sum_q y_{2q} \\ \vdots \\ \sum_q y_{kq} \end{pmatrix},$$

so the LS estimator of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} y_{1.} \\ y_{2.} \\ \vdots \\ y_{k.} \end{pmatrix},$$

where $y_{i.} = \sum_{q=1}^{n_i} y_{iq}/n_i$ is the mean(*) of the observations in the i th group. The estimator is in accordance with intuition since the observations are independent. We have

$$\mathbf{X}\hat{\boldsymbol{\theta}} = \begin{pmatrix} y_{1.} \\ \vdots \\ y_{1.} \\ y_{2.} \\ \vdots \\ y_{2.} \\ \vdots \\ y_{k.} \\ \vdots \\ y_{k.} \end{pmatrix} \begin{matrix} \left. \begin{matrix} y_{1.} \\ \vdots \\ y_{1.} \end{matrix} \right\} n_1 \text{ rows} \\ \left. \begin{matrix} y_{2.} \\ \vdots \\ y_{2.} \end{matrix} \right\} n_2 \text{ rows} \\ \vdots \\ \left. \begin{matrix} y_{k.} \\ \vdots \\ y_{k.} \end{matrix} \right\} n_k \text{ rows} \end{matrix}$$

(*) It should be observed that we are using the dot suffix on y to denote *averaging*, and not to denote summation as we did for frequencies and probabilities in Chapter 33.

and the SS due to the fitted model as a whole is

$$S_1 = (\mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{X}\hat{\boldsymbol{\theta}}) = \sum_{i=1}^k n_i y_i^2. \quad (35.16)$$

By subtraction, the Residual SS is, from (35.1),

$$S_R = \mathbf{y}'\mathbf{y} - S_1 = \sum_{i=1}^k \sum_{q=1}^{n_i} y_{iq}^2 - \sum_{i=1}^k n_i y_i^2 \equiv \sum_i \sum_q (y_{iq} - y_i.)^2. \quad (35.17)$$

To test the simple hypothesis imposing k constraints,

$$H_1: \boldsymbol{\theta} = \mathbf{0}, \quad (35.18)$$

we use (35.6) and obtain

$$F = \left(\frac{n-k}{k} \right) \frac{S_1}{S_R},$$

distributed as a $F'(k, n-k, \sum_i n_i \theta_i^2 / \sigma^2)$ variable, reducing to a (central) $F(k, n-k)$ variable if H_1 holds.

However, (35.18) is not the hypothesis of principal interest in most practical situations, where we usually wish to test whether the θ_i are all equal without specifying their common value. Instead of (35.18), we therefore test the composite hypothesis

$$H_2: \theta_1 - \theta_k = \theta_2 - \theta_k = \dots = \theta_{k-1} - \theta_k = 0, \quad (35.20)$$

which imposes only $(k-1)$ constraints. If (35.20) holds, the n observations are identically distributed with common mean

$$\theta_* \equiv \sum_{i=1}^k n_i \theta_i / n = \begin{pmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_k/n \end{pmatrix}' \boldsymbol{\theta}.$$

The LS estimator of θ_* is then the overall sample mean

$$\hat{\theta}_* = y_{..} = \frac{1}{n} \sum_{i=1}^k \sum_{q=1}^{n_i} y_{iq} = \frac{1}{n} \sum_i n_i y_i.$$

If $\mathbf{1}$ is a $(n \times 1)$ vector of units, we may rewrite the linear model temporarily in the (singular) form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \equiv \mathbf{1}\theta_* + \left[\mathbf{X} - \mathbf{1} \begin{pmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_k/n \end{pmatrix}' \right] \boldsymbol{\theta} + \boldsymbol{\epsilon}$$

and observe that the value of θ_* (a single constraint) is not involved in the hypothesis (35.20), and that the SS attributable to θ_* , namely

$$(\mathbf{1}\hat{\theta}_*)'(\mathbf{1}\hat{\theta}_*) = ny_{..}^2,$$

must be subtracted from (35.16) (the SS due to the fitted model as a whole) to give the SS due to the other $(k-1)$ constraints. This is

$$S_2 = S_1 - ny_{..}^2 \equiv \sum_{i=1}^k n_i (y_i. - y_{..})^2. \quad (35.21)$$

The Residual SS is given by S_R , defined at (35.17), as before.

35.8 now gives for the test statistic

$$F = \left(\frac{n-k}{k-1} \right) \frac{S_2}{S_R} \quad (35.22)$$

which is distributed as a $F'(k-1, n-k, \sum_{i=1}^k n_i (\theta_i - \theta_*)^2 / \sigma^2)$ variable, reducing to a central $F(k-1, n-k)$ when H_2 holds.

For computational purposes, S_2 and S_R are usually written as

$$\left. \begin{aligned} S_2 &= \sum_{i=1}^k \frac{\left(\sum_{q=1}^{n_i} y_{iq} \right)^2}{n_i} - \frac{\left(\sum_i \sum_q y_{iq} \right)^2}{n} \\ S_R &= \sum_i \sum_q y_{iq}^2 - \sum_i \frac{\left(\sum_q y_{iq} \right)^2}{n_i} \end{aligned} \right\} \quad (35.23)$$

and the results assembled in a table:

AV table for a one-way classification

Variation	D.fr.	SS	Mean square (MS) = SS/d. fr.
Between groups	$k-1$	S_2	$S_2/(k-1)$
Within groups	$n-k$	S_R	$S_R/(n-k)$
General mean	$n-1$ 1	$S_2 + S_R = \mathbf{y}' \mathbf{y} - ny_{..}^2$ $S_1 - S_2 = ny_{..}^2$	$ny_{..}^2$
TOTAL	n	$S_1 + S_R = \mathbf{y}' \mathbf{y}$	

(35.24)

The "General mean" row of (35.24) is generally omitted as of no interest; the variance-ratio test based on the ratio of $ny_{..}^2$ to $S_R/(n-k)$ is, of course, the ordinary "Student's" t^2 test for the mean, i.e. it has a $F(1, n-k)$ distribution when $\theta = \mathbf{0}$. The test (35.22) is simply the ratio of the "Between groups" MS to the "Within groups" MS, while (35.19) is obtained by adding together the "Between groups" and "General mean" rows of the table and taking the ratio of the resulting MS to the "Within groups" MS.

AV identities and their geometrical interpretations

35.12 The general theory of the linear model has been used in Example 35.1, but the final result can be less formally derived as follows. The identity

$$\sum_{i=1}^k \sum_{q=1}^{n_i} (y_{iq} - y_{..})^2 \equiv \sum_{i=1}^k \sum_{q=1}^{n_i} (y_{iq} - y_{i.})^2 + \sum_{i=1}^k n_i (y_{i.} - y_{..})^2 \quad (35.25)$$

splits the SS of the observations about their overall mean into a SS "within groups" and a SS "between groups" (i.e. between group means). If it can be verified that the two sums on the right of (35.25) are independently distributed in the χ'^2 form, the ratio of the second to the first is an intuitively acceptable criterion for testing the

equality of the group means in the population. This approach leads to (35.22) as before, but it offers no direct justification for the choice of this particular test statistic, for which the general theory is necessary. In more complicated situations, the approach through algebraic identities like (35.25) is often much simpler and quicker than the direct use of linear model theory, but care is necessary in splitting the SS—ultimately, safety lies only in checking with the general theory.

35.13 The Pythagorean form of (35.25) has the virtue of drawing attention to a geometrical interpretation of the algebraic partitioning of the SS which is the essence of AV. We saw in Example 11.7 that the simpler identity (for a single group of observations)

$$\sum_i y_i^2 \equiv \sum_i (y_i - \bar{y})^2 + n\bar{y}^2 \quad (35.26)$$

is geometrically equivalent to projecting the point $\mathbf{y} = (y_1, y_2, \dots, y_n)$ in the n -dimensional sample space upon the equiangular vector, which it meets at $(\bar{y}, \bar{y}, \dots, \bar{y})$, and using Pythagoras' theorem in the resulting right-angled triangle. In the more general notation which we have been using in Example 35.1 and in (35.25), (35.26) is

$$\sum_{i=1}^k \sum_{q=1}^{n_i} y_{iq}^2 \equiv \sum_i \sum_q (y_{iq} - y_{i..})^2 + n y_{..}^2, \quad (35.27)$$

and is therefore seen to be equivalent to the splitting-off of the "general mean" row from the Total SS in (35.24) to give the left-hand side of (35.25). The further decomposition in (35.25) of the first term on the right of (35.27) is similarly geometrically interpretable, $\sum_i \sum_q (y_{iq} - y_{i..})^2$ being the squared distance from \mathbf{y} to the vector $\mathbf{X}\hat{\boldsymbol{\theta}}$ defined above (35.16), and $\sum_i n_i (y_{i..} - y_{..})^2$ being the squared distance from $\mathbf{X}\hat{\boldsymbol{\theta}}$ to the equiangular vector. From the geometrical standpoint, therefore, AV is seen to consist of a resolution of the distance from \mathbf{y} to the origin into a number of components relevant to the problem in hand.

35.14 The fact that in Example 35.1 we obtained an orthogonal analysis for any classification into groups, no matter what the sizes n_i were, encourages us to investigate more complex classification systems. We shall find that orthogonality does not generally persist for unequal group-sizes, but does so when the sizes are equal. We first treat the case of a two-way classification in detail, since it exhibits most of the points of general interest.

AV in a two-way cross-classification

35.15 Suppose that, instead of being simply classified into k groups as in Example 35.1, a sample of n observations is classified in a $r \times c$ table with frequencies

n_{11}	n_{12}	\dots	n_{1c}	$n_{1.}$
n_{21}	n_{22}	\dots	n_{2c}	$n_{2.}$
\vdots	\vdots		\vdots	\vdots
n_{r1}	n_{r2}	\dots	n_{rc}	$n_{r.}$
$n_{.1}$	$n_{.2}$	\dots	$n_{.c}$	n

(35.28)

Although (35.28) is formally identical with (33.60), our present problem is distinguished from those of Chapter 33 by the fact that the value of y is here known for each observation, whereas there only the frequencies in the cells of the table were known. We express this distinction by referring here to a $r \times c$ *classification* as opposed to the term *categorization* used in Chapter 33. We continue the convention of Chapter 33 that a dot replacing a suffix to n denotes *summation* over that suffix, while for the variable y we continue the convention set up in Example 35.1 that a dot replacing a suffix denotes *averaging* over that suffix. Together, these two conventions simplify the notation in what follows. The reader will see that the grand total frequency in (35.28) should strictly be written $n_{..}$, but we continue to write n instead in this one case to denote "sample size."

We may, of course, treat the rc cells in the body of the table (35.28) as a one-way classification (Example 35.1) with $k = rc$. However, the questions which are usually asked about the two-way cross-classification (35.28) are:

- (1) Do the means of the row-classification (with frequencies $n_{1.}, n_{2.}, \dots, n_{r.}$) differ?
- (2) Do the means of the column-classification (with frequencies $n_{.1}, n_{.2}, \dots, n_{.c}$) differ?
- (3) Is there any interrelation between row- and column-means?

More rarely, we ask also

- (4) Does the mean of the whole set of n observations differ from some hypothetical value?

35.16 Denote the p th observation in the i th row and j th column of the table by y_{ijp} . We then have, in our notational convention,

$$\left. \begin{aligned} y_{ij.} &= \sum_{p=1}^{n_{ij}} y_{ijp} / n_{ij}, \\ y_{i..} &= \sum_{j=1}^c \sum_{p=1}^{n_{ij}} y_{ijp} / n_{i.} \equiv \sum_{j=1}^c n_{ij} y_{ij.} / n_{i.} \\ &\equiv \sum_{j=1}^c n_{ij} y_{ij.} / \sum_{j=1}^c n_{ij}, \\ y_{.j.} &= \sum_{i=1}^r \sum_{p=1}^{n_{ij}} y_{ijp} / n_{.j} \equiv \sum_{i=1}^r n_{ij} y_{ij.} / n_{.j} \\ &\equiv \sum_{i=1}^r n_{ij} y_{ij.} / \sum_{i=1}^r n_{ij}, \\ y_{...} &= \sum_{i=1}^r \sum_{j=1}^c \sum_{p=1}^{n_{ij}} y_{ijp} / n \equiv \sum_{i=1}^r n_{i.} y_{i..} / n \equiv \sum_{j=1}^c n_{.j} y_{.j.} / n. \end{aligned} \right\} \quad (35.29)$$

An easy way of avoiding any possible confusion in notation is to define a dummy variable n_{ijp} which is identically equal to 1 for all $p = 1, 2, \dots, n_{ij}$. Then (35.29) becomes

$$\left. \begin{aligned} y_{ij.} &= \sum_p y_{ijp} / \sum_p n_{ijp}, \\ y_{i..} &= \sum_j \sum_p y_{ijp} / \sum_j \sum_p n_{ijp}, \\ y_{.j.} &= \sum_i \sum_p y_{ijp} / \sum_i \sum_p n_{ijp}, \\ y_{...} &= \sum_i \sum_j \sum_p y_{ijp} / \sum_i \sum_j \sum_p n_{ijp}, \end{aligned} \right\} \quad (35.30)$$

which is easily remembered by its numerator-denominator symmetry. In (35.30), and hereafter unless otherwise stated, i is always summed from 1 to r ; j is summed from 1 to c ; and p is summed from 1 to n_{ij} .

35.17 In formulating the linear model, we require one parameter for the mean of the observations in each cell of the $r \times c$ table. In order to answer the questions posed in 35.15, however, we express the mean μ_{ij} in a cell in terms of:

- μ_{**} , a mean common to all observations;
- μ_{i*} , a mean common to all observations in the i th row;
- μ_{*j} , a mean common to all observations in the j th column.

Since we already have the cell means μ_{ij} , common to observations in the i th row and the j th column, we now have $1 + r + c + rc$ parameters, of which only rc can be linearly independent. The singularity which we have thus introduced by our choice of parameters can easily be removed, either by the augmentation technique of 19.14–15, Vol. 2, or by eliminating the redundant parameters, as we shall do here.

Once μ_{**} is defined, any $(r-1)$ of the means μ_{i*} determine the other one; similarly, only $(c-1)$ of the means μ_{*j} need be considered, since they with μ_{**} will determine the other one. Once the μ_{i*} and μ_{*j} are thus determined, it is easy to see that only $(r-1)(c-1)$ of the μ_{ij} can be independently determined (cf. the d.fr. in 33.29). We may thus confine ourselves to $(r-1)$ parameters μ_{i*} (omitting μ_{r*} , say), to $(c-1)$ parameters μ_{*j} (omitting μ_{*c} , say), and to $(r-1)(c-1)$ parameters μ_{ij} (say, $i = 1, 2, \dots, r-1$ and $j = 1, 2, \dots, c-1$). These, with μ_{**} , make up the rc parameters required for the model to be non-singular.

It should be noticed that we do not define the parameters μ_{**} , μ_{i*} , μ_{*j} except to state that they are (weighted) means of the μ_{ij} .

35.18 We now define

$$\left. \begin{aligned} \theta_{**} &= \mu_{**}, \\ \theta_{i*} &= \mu_{i*} - \mu_{**}, \\ \theta_{*j} &= \mu_{*j} - \mu_{**}, \\ \theta_{ij} &= \mu_{ij} - (\mu_{i*} + \mu_{*j}) + \mu_{**}, \end{aligned} \right\} \quad (35.31)$$

and write the linear model in the form

$$y_{ijp} = \theta_{**} + \theta_{i*} + \theta_{*j} + \theta_{ij} + \varepsilon_{ijp} \equiv \mu_{ij} + \varepsilon_{ijp}. \quad (35.32)$$

For obvious reasons, θ_{**} is called the *general mean*, and θ_{i*} , θ_{*j} are respectively called the *i th row-effect* and the *j th column-effect*, measuring the deviation from the general mean in a particular row or column. If the deviation of the cell-mean from the general mean were exactly equal to the sum of the corresponding row-effect and column-effect, we should have

$$\mu_{ij} - \mu_{**} = (\mu_{i*} - \mu_{**}) + (\mu_{*j} - \mu_{**}),$$

which implies $\theta_{ij} = 0$. We then say, in accordance with ordinary usage, that the i th

row and j th column "act additively" or "do not interact." θ_{ij} as defined in (35.31) measures departures from this situation, and is called the *interaction* between the i th row and the j th column.

35.19 The $(r+c+1)$ linear relations between the parameters, discussed in 35.17, may now be written (but we shall return to this subject in 35.26-8 below)

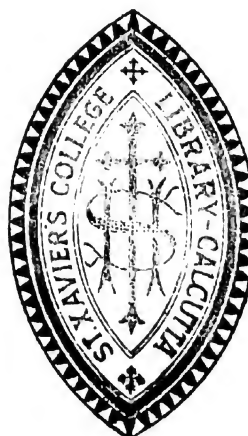
$$\left. \begin{aligned} 0 &= \sum_{i=1}^r n_{i.} \theta_{i.} = \sum_{j=1}^c n_{.j} \theta_{.j} \\ &= \sum_{i=1}^r n_{ij} \theta_{ij}, \quad j = 1, 2, \dots, c-1, \\ &= \sum_{j=1}^c n_{ij} \theta_{ij}, \quad i = 1, 2, \dots, r-1, \\ &= \sum_{i=1}^r \sum_{j=1}^c n_{ij} \theta_{ij}. \end{aligned} \right\} \quad (35.33)$$

If, as in 35.17, we define the parameters θ in (35.32) for $i = 1, 2, \dots, r-1$ and $j = 1, 2, \dots, c-1$ only, the eliminated $(r+c+1)$ parameters may be expressed in terms of the others, using (35.33), as

$$\left. \begin{aligned} \theta_{r.} &= - \sum_{i=1}^{r-1} n_{i.} \theta_{i.} / n_{r.}, \\ \theta_{.c} &= - \sum_{j=1}^{c-1} n_{.j} \theta_{.j} / n_{.c}, \\ \theta_{rj} &= - \sum_{i=1}^{r-1} n_{ij} \theta_{ij} / n_{rj}, \quad j = 1, 2, \dots, c-1, \\ \theta_{ic} &= - \sum_{j=1}^{c-1} n_{ij} \theta_{ij} / n_{ic}, \quad i = 1, 2, \dots, r-1, \\ \theta_{rc} &= + \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} \theta_{ij} / n_{rc}. \end{aligned} \right\} \quad (35.34)$$

35.20 We may now write down the matrix \mathbf{X} of the linear model (35.32). It is not a matrix of units and zeros only, because the expression of the eliminated parameters in terms of the others, in (35.34), involves various ratios of the n 's.

To simplify the reader's verification of the elements of the matrix in (35.35), its columns are headed by the parameters to which they correspond and its rows are bordered by the frequencies in the cells to which they apply. Only non-zero elements of \mathbf{X} are shown. Throughout the matrix, a vector of units $\mathbf{1}$ contains a number of components equal to the sum of the cell-frequencies (in the border of the rows) over which the vector $\mathbf{1}$ physically extends in (35.35).



$$X = \begin{pmatrix} n_{11} & n_{12} & \dots & n_{1,c-1} & n_{1c} \\ n_{21} & n_{22} & \dots & n_{2,c-1} & n_{2c} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ n_{r-1,1} & n_{r-1,2} & \dots & n_{r-1,c-1} & n_{r-1,c} \\ n_{r1} & n_{r2} & \dots & n_{r,c-1} & n_{rc} \end{pmatrix} = \begin{pmatrix} \theta_{**} & \theta_{1*} & \theta_{2*} & \dots & \theta_{r-1,*} & \theta_{*1} & \theta_{*2} & \dots & \theta_{*,c-1} & \theta_{11} & \theta_{12} & \dots & \theta_{1,c-1} \\ 1 & 1 & & & & 1 & & & & 1 & & & \\ & 1 & & & & & 1 & & & & 1 & & \\ & & \ddots & & & & & \ddots & & & & \ddots & \\ & & & 1 & & & & & 1 & & & & 1 \\ -\frac{n_{11}}{n_{1c}} & -\frac{n_{12}}{n_{1c}} & \dots & -\frac{n_{1,c-1}}{n_{1c}} & 1 & -\frac{n_{11}}{n_{1c}} & -\frac{n_{12}}{n_{1c}} & \dots & -\frac{n_{1,c-1}}{n_{1c}} & 1 & & & \\ 1 & 1 & & & & 1 & & & & & & & \\ & 1 & & & & & 1 & & & & & & \\ & & \ddots & & & & & \ddots & & & & \ddots & \\ & & & 1 & & & & & 1 & & & & 1 \\ -\frac{n_{21}}{n_{2c}} & -\frac{n_{22}}{n_{2c}} & \dots & -\frac{n_{2,c-1}}{n_{2c}} & 1 & & & & & & & & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \\ 1 & 1 & & & & 1 & & & & & & & \\ & 1 & & & & & 1 & & & & & & \\ & & \ddots & & & & & \ddots & & & & \ddots & \\ & & & 1 & & & & & 1 & & & & 1 \\ -\frac{n_{r-1,1}}{n_{r-1,c}} & -\frac{n_{r-1,2}}{n_{r-1,c}} & \dots & -\frac{n_{r-1,c-1}}{n_{r-1,c}} & 1 & & & & & & & & \\ 1 & 1 & & & & 1 & & & & & & & \\ & 1 & & & & & 1 & & & & & & \\ & & \ddots & & & & & \ddots & & & & \ddots & \\ & & & 1 & & & & & 1 & & & & 1 \\ -\frac{n_{11}}{n_{rc}} & -\frac{n_{12}}{n_{rc}} & \dots & -\frac{n_{1,c-1}}{n_{rc}} & 1 & -\frac{n_{11}}{n_{rc}} & -\frac{n_{12}}{n_{rc}} & \dots & -\frac{n_{1,c-1}}{n_{rc}} & 1 & & & \\ & 1 & & & & & 1 & & & & & & \\ & & \ddots & & & & & \ddots & & & & \ddots & \\ & & & 1 & & & & & 1 & & & & 1 \\ -\frac{n_{21}}{n_{rc}} & -\frac{n_{22}}{n_{rc}} & \dots & -\frac{n_{2,c-1}}{n_{rc}} & 1 & & & & & & & & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \\ 1 & 1 & & & & 1 & & & & & & & \\ & 1 & & & & & 1 & & & & & & \\ & & \ddots & & & & & \ddots & & & & \ddots & \\ & & & 1 & & & & & 1 & & & & 1 \end{pmatrix}$$

[illegible]

(35.35)



Premultiplying (35.35) by its transpose, we find that

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 1 & (r-1) & (c-1) & (r-1)(c-1) \\ n & 0' & 0' & 0' & 0' \\ 0 & \mathbf{A} & \mathbf{D} & \mathbf{0} & 0 \\ 0 & \mathbf{D}' & \mathbf{B} & \mathbf{0} & 0 \\ 0 & 0 & 0 & 0 & \mathbf{C} \end{pmatrix} \quad (35.36)$$

(The orders of the submatrices being indicated by the numbers bordering the matrix),
where \mathbf{A} , \mathbf{B} and \mathbf{C} are symmetric matrices with elements above the leading diagonal given by

$$\mathbf{A} = \begin{pmatrix} n_{1.} + \frac{n_{1.}^2}{n_r} & \frac{n_{1.}n_{2.}}{n_r} & \frac{n_{1.}n_{3.}}{n_r} & \dots & \frac{n_{1.}n_{r-1.}}{n_r} \\ & n_{2.} + \frac{n_{2.}^2}{n_r} & \frac{n_{2.}n_{3.}}{n_r} & \dots & \frac{n_{2.}n_{r-1.}}{n_r} \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & n_{r-1.} + \frac{n_{r-1.}^2}{n_r} \end{pmatrix}, \quad (35.37)$$

$$\mathbf{B} = \begin{pmatrix} n_{.1} + \frac{n_{.1}^2}{n_c} & \frac{n_{.1}n_{.2}}{n_c} & \frac{n_{.1}n_{.3}}{n_c} & \dots & \frac{n_{.1}n_{.c-1}}{n_c} \\ & n_{.2} + \frac{n_{.2}^2}{n_c} & \frac{n_{.2}n_{.3}}{n_c} & \dots & \frac{n_{.2}n_{.c-1}}{n_c} \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & n_{.c-1} + \frac{n_{.c-1}^2}{n_c} \end{pmatrix}. \quad (35.38)$$

The $(r-1)(c-1) \times (r-1)(c-1)$ matrix \mathbf{C} is more complicated. If we label its rows and columns by the suffixes of the θ_{ij} to which they refer (so that, for example, the 3rd row, c th column would be labelled the (13)th row, (21)th column) then the element

ANALYSIS OF VARIANCE IN THE LINEAR MODEL

in the (kl) th row and (mq) th column of \mathbf{C} is

$$C_{(kl), (mq)} = \left. \begin{aligned} & n_{kl} + n_{kl}^2 \left(\frac{1}{n_{kc}} + \frac{1}{n_{rl}} + \frac{1}{n_{rc}} \right) && \text{if } k = m, l = q, \\ & = n_{kl} n_{kq} \left(\frac{1}{n_{kc}} + \frac{1}{n_{rc}} \right) && \text{if } k = m, l \neq q, \\ & = n_{kl} n_{ml} \left(\frac{1}{n_{rl}} + \frac{1}{n_{rc}} \right) && \text{if } k \neq m, l = q, \\ & = n_{kl} n_{mq} / n_{rc} && \text{if } k \neq m, l \neq q. \end{aligned} \right\} \quad (35.39)$$

The only remaining non-null matrix in (35.36) is \mathbf{D} , of order $(r-1) \times (c-1)$, whose (i,j) th element is

$$D_{ij} = n_{ij} \left(1 - \frac{n_{ic}}{n_{\cdot c}} \frac{n_{\cdot j}}{n_{ij}} \right) + \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot c}} \left(\frac{n_{rc}}{n_{\cdot c}} - \frac{n_{rj}}{n_{\cdot j}} \right). \quad (35.40)$$

35.21 In general, $\mathbf{X}'\mathbf{X}$ at (35.36) can only be inverted numerically as in the general LS procedure, but inspection reveals that if we can make $\mathbf{D} = \mathbf{0}$, the matrix will be of the form

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \mathbf{A} & \mathbf{0} \\ & \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{B} & \mathbf{C} \end{pmatrix} \quad (35.41)$$

whose inverse is simply

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} n^{-1} & & \mathbf{0} \\ & \mathbf{A}^{-1} & \\ & & \mathbf{B}^{-1} \\ \mathbf{0} & & & \mathbf{C}^{-1} \end{pmatrix}. \quad (35.42)$$

We are therefore led to examine the conditions under which $\mathbf{D} = \mathbf{0}$, i.e. every element D_{ij} defined by (35.40) is zero. The structure of D_{ij} makes it evident that this will be so if and only if

$$\frac{n_{ij}}{n_{\cdot j}} = \frac{n_{ic}}{n_{\cdot c}}, \quad i = 1, 2, \dots, r-1; j = 1, 2, \dots, c-1,$$

and also

$$\frac{n_{rj}}{n_{\cdot j}} = \frac{n_{rc}}{n_{\cdot c}}.$$

These conditions are simply that every cell-frequency n_{ij} be proportional to its column-total frequency $n_{\cdot j}$. It follows that every n_{ij} must then also be proportional to its row-total frequency $n_{i\cdot}$, and that we must have

$$n_{ij} = n_{i\cdot} n_{\cdot j} / n, \quad \text{all } i, j, \quad (35.43)$$

for \mathbf{D} to be equal to the null matrix. Under this proportional frequencies condition, (*) the analysis of the two-way classification becomes relatively simple.

(*) The proportionality condition (35.43) is, the reader may recognize, precisely the condition determining the "independence" frequencies in a contingency table—cf. 33.4 and 33.29. The

The proportional-frequencies case

35.22 We first observe that the form of (35.42) implies that the LS estimator of the general mean $\theta_{..}$ is orthogonal to the estimators of all the other parameters, and similarly that the $(r-1)$ linearly independent row-effects are estimated orthogonally from all other parameters, as are the $(c-1)$ linearly independent column-effects and the $(r-1)(c-1)$ linearly independent interactions. The only non-orthogonalities occur *within* these last three groups of parameters, and have been imposed by the fact that each group has been obtained as a linearly independent subset of the larger (singular) group of parameters in which we are interested.

The reader will, perhaps, have observed that we have not yet evaluated the LS estimators themselves. The reason for this is that even when the proportional-frequencies condition (35.43) holds, the elements of \mathbf{C} given at (35.39) are not such as to make its inversion simple, although of course we may evaluate \mathbf{C}^{-1} numerically in any given situation. Fortunately, however, we may use the orthogonalities referred to in the preceding paragraph to obtain the LS estimators of the row- and column-effects at once, and use them later to evaluate the LS estimators of the interactions. To do this, we need only invert \mathbf{A} and \mathbf{B} at (35.37-8), and use (35.42) to evaluate the first $1+(r-1)+(c-1) = r+c-1$ rows of the $(rc \times 1)$ LS estimator vector.

35.23 It is easily verified by matrix multiplication that the inverse of (35.37) is

$$\mathbf{A}^{-1} = \begin{pmatrix} \frac{1}{n_{1.}} - \frac{1}{n}, & -\frac{1}{n}, & -\frac{1}{n}, & \dots, & -\frac{1}{n} \\ & \frac{1}{n_{2.}} - \frac{1}{n}, & -\frac{1}{n}, & \dots, & -\frac{1}{n} \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ & & & & \frac{1}{n_{r-1.}} - \frac{1}{n} \end{pmatrix}, \quad (35.44)$$

and similarly that (35.38) has inverse

$$\mathbf{B}^{-1} = \begin{pmatrix} \frac{1}{n_{.1}} - \frac{1}{n}, & -\frac{1}{n}, & -\frac{1}{n}, & \dots, & -\frac{1}{n} \\ & \frac{1}{n_{.2}} - \frac{1}{n}, & -\frac{1}{n}, & \dots, & -\frac{1}{n} \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ & & & & \frac{1}{n_{.c-1}} - \frac{1}{n} \end{pmatrix}. \quad (35.45)$$

resemblance is merely formal, since in Chapter 33 the n_{ij} were random variables, whereas here they are predetermined constants.

We now require only the first $r+c-1$ rows of the $(rc \times 1)$ vector $\mathbf{X}'\mathbf{y}$. From (35.35), these are seen to be, in the notation of (35.29),

$$(\mathbf{X}'\mathbf{y})_{r+c-1} = \begin{pmatrix} ny_{...} \\ n_{1.}(y_{1..} - y_{r..}) \\ n_{2.}(y_{2..} - y_{r..}) \\ \vdots \\ n_{r-1.}(y_{r-1..} - y_{r..}) \\ n_{.1}(y_{.1.} - y_{.c.}) \\ n_{.2}(y_{.2.} - y_{.c.}) \\ \vdots \\ n_{.c-1}(y_{.c-1.} - y_{.c.}) \end{pmatrix}. \quad (35.46)$$

Using (35.42) and (35.44-6) we find, for the first $(r+c-1)$ components of the LS estimator $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$,

$$\begin{pmatrix} \hat{\theta}_{**} \\ \hat{\theta}_{1*} \\ \vdots \\ \hat{\theta}_{r-1,*} \\ \hat{\theta}_{*1} \\ \vdots \\ \hat{\theta}_{*c-1} \end{pmatrix} = \begin{pmatrix} y_{...} \\ y_{1..} - y_{r..} \\ \vdots \\ y_{r-1..} - y_{r..} \\ y_{.1.} - y_{.c.} \\ \vdots \\ y_{.c-1.} - y_{.c.} \end{pmatrix}. \quad (35.47)$$

Thus the LS estimator of the general mean is the overall sample mean, and the LS estimators of row- (or column-) effects are the sample differences between row- (or column-) means and the overall sample mean. It follows at once from (35.47) and the first two linear relationships in (35.34) that the same holds true for the eliminated (redundant) row- and column-effects, i.e. that

$$\hat{\theta}_{r*} = y_{r..} - y_{...}, \quad \hat{\theta}_{*c} = y_{.c.} - y_{...}. \quad (35.48)$$

35.24 Substituting (35.47-8) into the definition of the interactions θ_{ij} in (35.31), we see that

$$\hat{\theta}_{ij} = \hat{\mu}_{ij} - y_{i..} - y_{.j.} + y_{...}, \quad (35.49)$$

since θ_{ij} is a linear function of the other quantities (cf. 19.6, Vol. 2). Now, clearly, from the extreme right of (35.32), we must have the LS estimator

$$\hat{\mu}_{ij} = y_{ij.},$$

and thus (35.49) becomes

$$\hat{\theta}_{ij} = y_{ij.} - y_{i..} - y_{.j.} + y_{...}. \quad (35.50)$$

Thus all the parameters are estimated, in this proportional-frequencies case, by the "obvious" intuitive estimators.

Now that the LS estimators of all the parameters in our model are known, we may proceed, in Example 35.2, to test the various hypotheses corresponding to the questions asked in 35.15.

Example 35.2 Two-way cross-classification with proportional frequencies

The results of our investigations so far show that the linear model (35.32) for the

two-way cross-classification is representable in the non-singular form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad (35.51)$$

where \mathbf{X} is defined by (35.35), and $\boldsymbol{\theta}$ is the vector whose transpose heads the rows of \mathbf{X} in (35.35).

We first consider question (1) in 35.15. This corresponds to asking for a test of the hypothesis

$$H_1: \theta_{1*} = \theta_{2*} = \dots = \theta_{r-1,*} = 0, \quad (35.52)$$

imposing $(r-1)$ constraints. Question (2) in 35.15 is similarly equivalent to the $(c-1)$ -constraint hypothesis

$$H_2: \theta_{*1} = \theta_{*2} = \dots = \theta_{*c-1} = 0, \quad (35.53)$$

and question (3) in 35.15 to the $(r-1)(c-1)$ -constraint hypothesis

$$H_3: \theta_{ij} = 0, \quad i = 1, 2, \dots, r-1; j = 1, 2, \dots, c-1. \quad (35.54)$$

Finally, question (4) in 35.15 corresponds to the single-constraint hypothesis

$$H_4: \theta_{**} = 0. \quad (35.55)$$

All four hypotheses (35.52-5) are composite. To test any one of them, we must find the SS attributable to that hypothesis, and use the general theory which we have developed, summarized in 35.8. Since the four hypotheses between them account for all rc parameters in the model, and have no parameter in common, we see that we have to partition the SS due to the fitted model as a whole, namely $\hat{\boldsymbol{\theta}}' \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\theta}}$, into the components attributable to the four hypotheses. This is particularly straightforward here, since we have seen in 35.22 that the four groups of parameters are estimated by orthogonal sets of estimators. In fact, $\mathbf{X}' \mathbf{X}$ was given at (35.41).

We now write the LS estimators from (35.47) and (35.50) in the form

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} y_{...} \\ \vdots \\ y_{1..} - y_{...} \\ \vdots \\ y_{r-1..} - y_{...} \\ \vdots \\ y_{.1.} - y_{...} \\ \vdots \\ y_{.c-1.} - y_{...} \\ \vdots \\ y_{11.} - y_{1..} - y_{.1.} + y_{...} \\ \vdots \\ y_{r-1,c-1.} - y_{r-1..} - y_{.c-1.} + y_{...} \end{pmatrix} \equiv \begin{pmatrix} \hat{\theta}_{**} \\ \vdots \\ \hat{\theta}_{i*} \\ \vdots \\ \hat{\theta}_{*j} \\ \vdots \\ \hat{\theta}_{ij} \end{pmatrix}, \quad (35.56)$$

where the subvectors of $\hat{\boldsymbol{\theta}}$ have $1, r-1, c-1$ and $(r-1)(c-1)$ components respectively. From (35.56) and (35.41), we have the decomposition

$$\hat{\boldsymbol{\theta}}' \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\theta}} = n\hat{\theta}_{**}^2 + \hat{\boldsymbol{\theta}}_{i*}' \mathbf{A} \hat{\boldsymbol{\theta}}_{i*} + \hat{\boldsymbol{\theta}}_{*j}' \mathbf{B} \hat{\boldsymbol{\theta}}_{*j} + \hat{\boldsymbol{\theta}}_{ij}' \mathbf{C} \hat{\boldsymbol{\theta}}_{ij}, \quad (35.57)$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are the submatrices of (35.41) defined at (35.37-9). The first term on the right of (35.57) is the SS attributable to H_4 , which we write explicitly as

$$S_4 = ny_{...}^2. \quad (35.58)$$

(35.37) may be written in the form

$$\mathbf{A} = \mathbf{D}_{n_i} + \mathbf{n}_r \mathbf{n}_r'$$

where \mathbf{D}_{n_i} is a $(r-1) \times (r-1)$ diagonal matrix with elements $n_{i..}$ and \mathbf{n}_r is the $(r-1) \times 1$ vector with elements $\frac{n_{i..}}{n_r^{1/2}}$. The second term on the right of (35.57) is now seen to be

$$\begin{aligned} S_1 &= \hat{\theta}_{i*}' \{ \mathbf{D}_{n_i} + \mathbf{n}_r \mathbf{n}_r' \} \hat{\theta}_{i*} \\ &= \hat{\theta}_{i*}' \mathbf{D}_{n_i} \hat{\theta}_{i*} + \hat{\theta}_{i*}' \mathbf{n}_r (\hat{\theta}_{i*}' \mathbf{n}_r)' \\ &= \sum_{i=1}^{r-1} n_{i..} (y_{i..} - y_{...})^2 + \left\{ \sum_{i=1}^{r-1} \frac{n_{i..}}{n_r^{1/2}} (y_{i..} - y_{...}) \right\}^2 \\ &= \sum_{i=1}^r n_{i..} (y_{i..} - y_{...})^2. \end{aligned} \quad (35.59)$$

This is the SS attributable to H_1 . In an exactly similar way, using \mathbf{B} at (35.38), we find for the third term on the right of (35.57)

$$S_2 = \sum_{j=1}^c n_{.j} (y_{.j} - y_{...})^2, \quad (35.60)$$

the SS attributable to H_2 . Finally, we find from (35.39) that the last term on the right of (35.57) is

$$\begin{aligned} S_3 &= \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} (y_{ij.} - y_{i..} - y_{.j} + y_{...})^2 + \frac{1}{n_{rc}} \left\{ \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} (y_{ij.} - y_{i..} - y_{.j} + y_{...}) \right\}^2 \\ &\quad + \sum_{i=1}^{r-1} \frac{1}{n_{ic}} \left\{ \sum_{j=1}^{c-1} n_{ij} (y_{ij.} - y_{i..} - y_{.j} + y_{...}) \right\}^2 \\ &\quad + \sum_{j=1}^{c-1} \frac{1}{n_{rj}} \left\{ \sum_{i=1}^{r-1} n_{ij} (y_{ij.} - y_{i..} - y_{.j} + y_{...}) \right\}^2 \\ &= \sum_{i=1}^r \sum_{j=1}^c n_{ij} (y_{ij.} - y_{i..} - y_{.j} + y_{...})^2, \end{aligned} \quad (35.61)$$

upon use of linear relationships among the estimated interactions precisely analogous to those for the interaction parameters given in (35.34). The four SS defined in (35.58-61) exhaust the SS due to the fitted model (35.51). The only other quantity we shall require is the Residual SS, which here, as generally, is the difference

$$\begin{aligned} S_R &= \mathbf{y}' \mathbf{y} - \hat{\theta}' \mathbf{X}' \mathbf{X} \hat{\theta} \\ &= \sum_i \sum_j \sum_p y_{ijp}^2 - (S_1 + S_2 + S_3 + S_4). \end{aligned} \quad (35.62)$$

For computational purposes, the other SS are written in the forms

$$\left. \begin{aligned} S_4 &= (\sum_i \sum_j \sum_p y_{ijp})^2 / n, \\ S_1 &= \sum_i \{ (\sum_j \sum_p y_{ijp})^2 / n_{i.} \} - (\sum_i \sum_j \sum_p y_{ijp})^2 / n, \\ S_2 &= \sum_j \{ (\sum_i \sum_p y_{ijp})^2 / n_{.j} \} - (\sum_i \sum_j \sum_p y_{ijp})^2 / n, \\ S_3 &= \sum_i \sum_j \{ (\sum_p y_{ijp})^2 / n_{ij} \} - \sum_i \{ (\sum_j \sum_p y_{ijp})^2 / n_{i.} \} \\ &\quad - \sum_j \{ (\sum_i \sum_p y_{ijp})^2 / n_{.j} \} + (\sum_i \sum_j \sum_p y_{ijp})^2 / n, \end{aligned} \right\} \quad (35.63)$$

which the reader should verify.

On substituting from (35.63), (35.62) becomes simply

$$S_R = \sum_i \sum_j \sum_p y_{ijp}^2 - \sum_i \sum_j \{(\sum_p y_{ijp})^2 / n_{ij}\}. \quad (35.64)$$

We may now, as we did in Example 35.1, assemble the results of our analysis in a table:

AV table for a two-way cross-classification with proportional frequencies

Variation	D.fr.	SS	MS	Non-central parameter λ
Between rows	$r-1$	S_1	$S_1/(r-1)$	$\sum_i n_{i.} \theta_{i.}^2 / \sigma^2$
Between columns	$c-1$	S_2	$S_2/(c-1)$	$\sum_j n_{.j} \theta_{.j}^2 / \sigma^2$
Interactions	$(r-1)(c-1)$	S_3	$S_3/(r-1)(c-1)$	$\sum_i \sum_j n_{ij} \theta_{ij}^2 / \sigma^2$
Residual	$\frac{n-rc}{n-1}$	S_R	$S_R/(n-rc)$	
General mean	1	S_4	S_4	$n\theta_{**}^2 / \sigma^2$
TOTAL	n	$y'y$		

(35.65)

The general theory of 35.8 tells us that the LR test of any of the hypotheses H_1 to H_4 is obtained by using the ratio of the corresponding MS in (35.65) to the Residual MS, and rejecting the hypothesis for large values of the ratio. Each of these ratios is a non-central F variate with d.fr. as given in the table and non-central parameter (obtained by using the general rule in 35.8) given in the last column of the table. (As in Example 35.1, the test for the general mean is the usual "Student's" t^2 -test.)

To test the comprehensive hypothesis

$$H_0: \theta = 0 \quad (35.66)$$

for all the parameters (which means that H_1, H_2, H_3 and H_4 all hold), the same theory tells us that the ratio to be used is

$$F = \frac{(S_1 + S_2 + S_3 + S_4)/rc}{S_R/(n-rc)}, \quad (35.67)$$

which is a F' $\{rc, n-rc, \sum_i \sum_j n_{ij}(\theta_{**} + \theta_{i.} + \theta_{.j} + \theta_{ij})^2 / \sigma^2\}$ variable, the non-central parameter being obtained from the last term on the right of (35.64), by substituting θ for $\hat{\theta}$ in accordance with the general rule. This test is exactly the one mentioned in 35.15, in which the rc cell-frequencies are treated as a one-way classification and the test (35.19) applied. Similarly, to test that H_1, H_2 and H_3 (but not H_4) hold, the numerator of (35.67) is replaced by $(S_1 + S_2 + S_3)/(rc-1)$, this test being equivalent to (35.22) applied to the rc cell-frequencies.

The equal-frequencies (balanced) case

35.25 The most important case of the proportional-frequencies situation (35.43) arises when all cell-frequencies n_{ij} are equal, say to m . The arithmetic of the computing formulae (35.63–4) then simplifies obviously (cf. Exercise 35.1). The matrix \mathbf{C} of (35.39) also now becomes easy to invert and the theory of 35.22–4 correspondingly more direct (cf. Exercise 35.2).

Apart from these simplifications, the only new point arising in this *balanced* case occurs when $m = 1$, for then (cf. Exercise 35.1) the Residual SS (35.64) is identically zero, as are its d.fr., $(n - rc)$. Since all the tests given in Example 35.2 then become nugatory, this situation clearly requires special consideration. It is not difficult to see how our problem comes about, for with $m = 1$, $n = rc$ we are put in the position of having to estimate rc parameters from the same number of observations. Not surprisingly, we can do this exactly, with no residual variation—we are in just the same position as we should be in fitting a polynomial of degree $q - 1$ (requiring q constants) to a set of q observations. Thus we can estimate all rc parameters even when $m = 1$, but only at the expense of seeing our Residual SS disappear.

There is no way out of this difficulty unless we consent to reduce the number of parameters in the model, and what we shall in fact do is to discard the $(r - 1)(c - 1)$ interaction parameters θ_{ij} , leaving ourselves with $r + c - 1$ parameters to be estimated. We shall then have a new Residual SS to replace (35.64), and in fact this will be seen in Example 35.3 to be precisely the former Interaction SS, S_3 , defined at (35.61).

It should not need to be stressed that this restricted model, without interaction parameters, is unsuitable for the analysis of data where interactions do exist. For this reason it is inadvisable to restrict ourselves voluntarily to one observation per cell of a cross-classification unless we are sure that rows and columns do not interact. However, considerations of cost or time sometimes enforce such a restriction.

Example 35.3 Two-way cross-classification with exactly one observation per cell

If the interaction parameters θ_{ij} are dropped from the linear model, we now have, with one observation per cell,

$$y_{ij} = \theta_{**} + \theta_{i*} + \theta_{*j} + \varepsilon_{ij},$$

where, to avoid singularities, we define θ_{i*} for $i = 1, 2, \dots, r - 1$ and θ_{*j} for $j = 1, 2, \dots, c - 1$, as previously. All the work of 35.17–19 in respect of our present parameters holds good. The matrix \mathbf{X} defined at (35.35) remains valid if we use only its first $(r + c - 1)$ columns, as does the leading $(r + c - 1) \times (r + c - 1)$ submatrix of $\mathbf{X}'\mathbf{X}$ at (35.36), in which we now still have $\mathbf{D} = \mathbf{0}$ since the proportional-frequencies condition (35.43) holds here. \mathbf{A} and \mathbf{B} at (35.37–8) and their inverses at (35.44–5) are unaffected, as are the vectors (35.46–7), which are now complete instead of partial vectors for the LS estimators of our parameters. We may therefore test the hypotheses H_1 , H_2 and H_4 at (35.52–3) and (35.55) exactly as in Example 35.2, the only difference being that what was previously the Interaction SS, S_3 , now becomes the new Residual SS, for the four SS in the following abbreviated table must add to $\mathbf{y}'\mathbf{y}$, as always.

AV table for a two-way cross-classification with one observation per cell

Variation	D.fr.	SS
Between rows	$r-1$	S_1
Between columns	$c-1$	S_2
Residual	$(r-1)(c-1)$	S_3
General mean	$rc-1$	S_4
	1	
TOTAL	n	$y'y$

(35.68)

The tests of H_1 , H_2 and H_4 can now be carried out with the MS $S_3/(r-1)(c-1)$ as denominator of the F statistic.

Although we have dropped the interaction parameters θ_{ij} from the model in order to obtain a Residual SS, we can also use their estimators $\hat{\theta}_{ij}$ to test for zero interactions by separating off from that Residual SS an appropriate component.

Consider the linear form

$$L = \sum_i \sum_j c_{ij} \hat{\theta}_{ij},$$

where $\hat{\theta}_{ij}$ is defined by (35.50) (the final suffix to the y 's is now redundant) and the c_{ij} are coefficients to be determined. If the interactions θ_{ij} are all zero, but in general not otherwise, $E(L) = 0$, and it is thus intuitively reasonable to use a statistic of the form L to test the hypothesis of zero interactions. If we choose the c_{ij} so that $\sum_i c_{ij} = \sum_j c_{ij} = 0$, we see from (35.50) that

$$L = \sum_i \sum_j c_{ij} y_{ij},$$

and hence

$$\text{var } L = C^2 \sigma^2,$$

where $C^2 = \sum_i \sum_j c_{ij}^2$, and σ^2 is the error variance as usual. Thus $L^2/(C^2 \sigma^2)$ is a χ^2 variable with 1 d.fr. when the interactions are zero. Moreover, our present Residual SS at (35.61) is $S_3 = \sum_i \sum_j \hat{\theta}_{ij}^2$, and $S_3 - \{L^2/(C^2 \sigma^2)\}$ is independent of $L^2/(C^2 \sigma^2)$, since the θ_{ij} can be orthogonally transformed to a set of standardized independent normal variates of which one is $L/C\sigma$, and $S_3 - \{L^2/(C^2 \sigma^2)\}$ will be the sum of squares of the others, distributed as χ^2 with $(r-1)(c-1) - 1 = rc - r - c$ d.fr.

It remains to choose the c_{ij} . They can be functions of the $\hat{\theta}_{i*}$, $\hat{\theta}_{*j}$, since the latter are distributed independently of the $\hat{\theta}_{ij}$ by (35.42), and hence the marginal distribution of L will be the same as any of its conditional distributions for fixed $\hat{\theta}_{i*}$, $\hat{\theta}_{*j}$, which will be as given above.

A simple choice is $c_{ij} = \hat{\theta}_{i*} \hat{\theta}_{*j}$, so that we may define

$$\begin{aligned} S_I &= (\sum_i \sum_j \hat{\theta}_{i*} \hat{\theta}_{*j} y_{ij})^2 / (\sum_i \hat{\theta}_{i*}^2 \sum_j \hat{\theta}_{*j}^2) \\ &= \frac{\sum_i \sum_j \{(y_{i.} - y_{..})(y_{.j} - y_{..}) y_{ij}\}^2}{\sum_i (y_{i.} - y_{..})^2 \sum_j (y_{.j} - y_{..})^2} \end{aligned}$$

(35.69)

S_1/σ^2 is a χ^2 variable with 1 d.fr. and $(S_3 - S_1)/\sigma^2$ independently a χ^2 variable with $(rc - r - c)$ d.fr. Their ratio $S_1/(S_3 - S_1) = F$ has the variance-ratio distribution with $(1, rc - r - c)$ d.fr., and may be used to test the hypothesis that all interactions are zero. This test for complete *additivity* of effects was suggested by Tukey (1949), who generalized it further—see Scheffé (1959). M. N. Ghosh and Sharma (1963) studied its power against the alternative that there are interactions of form $\theta_{ij} = \alpha\theta_{i*}\theta_{*j}$. For the 6×6 classification, the power was found to be of the same order as the F -test for interactions obtained by equating adjacent pairs of the θ_{i*} and of the θ_{*j} .

Choice of weights

35.26 We must now discuss a point which we deliberately passed over in formulating our linear model in 35.17–19. We observed there that we had $(r + c + 1)$ parameters in our original model which were redundant in the sense that they were linearly dependent upon the rc other parameters, and we therefore eliminated them using the set of linear relations given in (35.33), leading to (35.34), which determined the structure of the basic matrix (35.35). It is now necessary to recognize that the set of relations given in (35.33) is essentially arbitrary—in the first relation given there, for example, we chose to equate to zero the particular linear function $\sum_i n_i \theta_{i*}$, using as weights the marginal row frequencies n_i . This may seem natural, but it is by no means necessary: we might have chosen instead to use equal weights, so that $\sum_i \theta_{i*} = 0$, or indeed any weights w_i , so that $\sum_i w_i \theta_{i*} = 0$.

If the complete set of n observations were a simple random sample from some population, the observed n_i/n would be estimates of the population relative frequencies in the row categories, and it would therefore be meaningful to define the row-effects using these weights to express their linear dependence. Similarly, $\sum_j n_j \theta_{*j} = 0$, and the other relations in (35.33) would be meaningful in the same context. We call these the *frequency weights*.

35.27 However, in many experimental contexts there is no question of the observations being a random sample from some population—the $r \times c$ cross-classification is deliberately set up to throw light on the variable (y) being studied. The use of observed frequencies as weights in the linear relations (35.33) is then no longer readily interpretable. It may even be meaningless to consider any set of weights as the “right” ones, in the sense of reflecting an underlying population distribution; for example, if we have a 2×3 cross-classification to study the effects of two different doses of Fertilizer A and three different doses of Fertilizer B on the yield of a crop (y), one may be simply interested in the effects and interactions as such, and not as representing any population at all. There is a crucial distinction here between the “experimental” and the “survey” approach to data, to which we shall revert in Chapters 38–9.

In experimental investigations, therefore, it is common (for lack of any known appropriate system of weights) to use equal weights throughout (35.33). For the remainder of this chapter, the *equal-weights* system means that (35.33) holds with all symbols n_i, n_j, n_{ij} suppressed, i.e. replaced by 1's. It is to be observed that, whereas

in the balanced case (cf. 35.25) the equal-weights system has in effect already been used, simply because the frequencies were equal, our general results for the proportional-frequencies case (in 35.22-4 and Example 35.2) would not hold unless the frequency weights were used.

Now that we are about to resume discussion of the general disproportional-frequencies case, which we left in 35.21, the distinction between these two weighting systems will become acute, if only because, in this most general case, we may have a very small number of very large frequencies which tend to dominate the frequency weighting system, and perhaps distort its interpretation.

35.28 The choice of weights in (35.33) will in general affect the estimation of all the parameters, general mean, row- and column-effects, and interactions. However, if the true interactions θ_{ij} are all zero under any weighting system, they will be so for all weighting systems, as Scheffé (1959) proves. For under the first weighting system, (35.31) shows that $\theta_{ij}^{(1)} = 0$ is equivalent to

$$\mu_{ij} = \mu_{i*}^{(1)} + \mu_{*j}^{(1)} - \mu_{**}^{(1)}.$$

Under any other weighting system, the interactions are, from (35.31),

$$\begin{aligned}\theta_{ij}^{(2)} &= \mu_{ij} - \mu_{i*}^{(2)} - \mu_{*j}^{(2)} + \mu_{**}^{(2)} \\ &= (\mu_{i*}^{(1)} + \mu_{*j}^{(1)} - \mu_{**}^{(1)}) - (\mu_{i*}^{(2)} + \mu_{*j}^{(2)} - \mu_{**}^{(2)}) \\ &= (\mu_{i*}^{(1)} - \mu_{i*}^{(2)}) + (\mu_{*j}^{(1)} - \mu_{*j}^{(2)}) - (\mu_{**}^{(1)} - \mu_{**}^{(2)}).\end{aligned}$$

This is of the form

$$\theta_{ij}^{(2)} = a_i + b_j + c,$$

and it is evident from the definition of interactions in 35.18 that if they were representable as the sum of row-, column- and general components, these would be absorbed into the row-effects, column-effects and the general mean respectively, leaving the interaction equal to zero. We thus have $\theta_{ij}^{(2)} = 0$ for all i, j . If H_3 of (35.54) holds, therefore, it holds for every weighting system.

Disproportional frequencies

35.29 We first use the frequency weights (35.33) as before. The proportionality condition (35.43) does not hold, so that the matrix \mathbf{D} in (35.36) is non-null, and the simplified analysis of 35.22-5 is no longer valid. It remains true even in this most general case that (35.36) may be partitioned into

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & & \\ & \begin{pmatrix} \mathbf{A} & \mathbf{D} \\ \mathbf{D}' & \mathbf{B} \end{pmatrix} & \\ & & \mathbf{C} \end{pmatrix},$$

of which (35.41) was the special case $\mathbf{D} = 0$. The inverse of this can still be written down, for

$$\mathbf{E} = \begin{pmatrix} \mathbf{A} & \mathbf{D} \\ \mathbf{D}' & \mathbf{B} \end{pmatrix}^{-1} = \begin{pmatrix} (\mathbf{A} - \mathbf{D}\mathbf{B}^{-1}\mathbf{D}')^{-1} & -(\mathbf{A} - \mathbf{D}\mathbf{B}^{-1}\mathbf{D}')^{-1}\mathbf{D}\mathbf{B}^{-1} \\ -(\mathbf{B} - \mathbf{D}'\mathbf{A}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{A}^{-1} & (\mathbf{B} - \mathbf{D}'\mathbf{A}^{-1}\mathbf{D})^{-1} \end{pmatrix}, \quad (35.70)$$

as may be verified by multiplication, so that

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} n^{-1} & & \\ & \mathbf{E} & \\ & & \mathbf{C}^{-1} \end{pmatrix}. \quad (35.71)$$

The effect of the non-nullity of \mathbf{D} on the LS analysis is to change the estimator of the parameter-vector $\boldsymbol{\theta}$, for although the partial vector (35.46) is unchanged the $(r+c-1) \times (r+c-1)$ leading diagonal submatrix of (35.42) is now replaced by that of (35.71). If we write (35.46) concisely as

$$\begin{pmatrix} ny... \\ \mathbf{v}_{r-1} \\ \mathbf{v}_{c-1} \end{pmatrix},$$

each \mathbf{v} being the subvector of (35.46) with number of rows indicated by its suffix, we may generalize (35.47), using (35.70-1), to

$$\begin{aligned} ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y})_{r+c-1} &= \begin{pmatrix} n^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{E} \end{pmatrix} \begin{pmatrix} ny... \\ \mathbf{v}_{r-1} \\ \mathbf{v}_{c-1} \end{pmatrix} \\ &= \begin{pmatrix} y... \\ (\mathbf{A} - \mathbf{D}\mathbf{B}^{-1}\mathbf{D}')^{-1}(\mathbf{v}_{r-1} - \mathbf{D}\mathbf{B}^{-1}\mathbf{v}_{c-1}) \\ (\mathbf{B} - \mathbf{D}'\mathbf{A}^{-1}\mathbf{D})^{-1}(\mathbf{v}_{c-1} - \mathbf{D}'\mathbf{A}^{-1}\mathbf{v}_{r-1}) \end{pmatrix}. \end{aligned} \quad (35.72)$$

Thus the estimators $\hat{\theta}_{i*}$, $\hat{\theta}_{*j}$ are numerically determinable, while $\hat{\theta}_{**} = y...$ always, as is intuitively obvious. As in 35.24, the definition of the interactions at (35.31) then implies that their estimators satisfy

$$\hat{\theta}_{ij} = y_{ij} - \hat{\theta}_{i*} - \hat{\theta}_{*j} - y..., \quad (35.73)$$

so that the LS estimators of all the parameters are determined. The generalization of the decomposition (35.57) is

$$\hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}} = n\hat{\theta}_{**}^2 + \begin{pmatrix} \hat{\boldsymbol{\theta}}_{i*} \\ \hat{\boldsymbol{\theta}}_{*j} \end{pmatrix}' \begin{pmatrix} \mathbf{A} & \mathbf{D} \\ \mathbf{D}' & \mathbf{B} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\theta}}_{i*} \\ \hat{\boldsymbol{\theta}}_{*j} \end{pmatrix} + \hat{\boldsymbol{\theta}}_{ij}'\mathbf{C}\hat{\boldsymbol{\theta}}_{ij}. \quad (35.74)$$

Example 35.4 Two-way cross-classification with disproportional frequencies and frequency weights

(35.74) shows at once that H_3 and H_4 of (35.54-5) can each be tested in the manner of Example 35.2, although the SS attributable to the interactions must now be numerically evaluated from the last term on the right of (35.74). Thus both the general mean and the interactions MS can be tested (with 1 and $(r-1)(c-1)$ d.fr. respectively) against the Residual MS, since they are non-central F variables, irrespective of the row- and column-effects. (*)

The SS attributable to the row- and column-effects jointly is the middle term on the

(*) It should be particularly noted that if equal weights were used instead of frequency weights in (35.33), the SS attributable to interactions would not be a separate component in (35.74) but would be entangled with that for row- and column-effects just as the latter are entangled with each other. However, H_3 would hold, if true, whichever weighting system were used, in virtue of the result of 35.28.

right of (35.74), say M . H_1 and H_2 of (35.52-3) could therefore be tested *jointly* by calculating M . However, practical interest usually lies in testing row- and column-effects separately. The SS attributable to row-effects, for example, would be obtained by calculating the reduction in the Residual SS brought about by first estimating all parameters except θ_{i*} and then estimating all parameters including θ_{i*} —this is the SS attributable to θ_{i*} . Similarly, θ_{*j} can have its SS evaluated. These two SS will not add to M , since row- and column-effects are not orthogonal in general.

If it can be assumed that all interactions are zero, the situation simplifies (cf. Exercise 35.4).

35.30 Use of the equal-weights system instead of (35.33) makes the testing of row- and of column-effects computationally a good deal simpler than with the frequency weights used in Example 35.4. We may proceed directly as follows.

Suppose that we first analyse the $r \times c$ cross-classification as if it were a one-way classification with $k = rc$. We then obtain an AV with $n - rc$ d.fr. for Residual, the remaining rc d.fr. being attributable to the combined effect of row- and column-classifications. Using Example 35.1, we find the AV table below:

AV for any $r \times c$ cross-classification

Variation	D.fr.	SS	
Due to classification as a whole	rc	$\sum_i \sum_j n_{ij} y_{ij}^2$	(35.75)
Residual	$n - rc$	$\sum_i \sum_j \sum_p (y_{ijp} - y_{ij.})^2$	
TOTAL	n	$\sum_i \sum_j \sum_p y_{ijp}^2$	

It is clear that any cell for which $n_{ij} = 1$ can contribute nothing to the Residual MS (cf. 35.25 and Example 35.3 where *every* cell has $n_{ij} = 1$). To split the rc d.fr. for the classification into its component parts due to row-effects, column-effects, interactions and the general mean, we need only analyse the cell means $y_{ij.}$.

35.31 From the model (35.32), it follows that the $y_{ij.}$ satisfy

$$y_{ij.} = \theta_{**} + \theta_{i*} + \theta_{*j} + \theta_{ij} + \varepsilon_{ij}, \quad (35.76)$$

where the errors ε_{ij} are uncorrelated, with zero means and variances σ^2/n_{ij} . Suppose now that we average the $y_{ij.}$ over columns, i.e. take the *unweighted* mean $\frac{1}{c} \sum_{j=1}^c y_{ij.} = \bar{y}_i$, say. It then follows from (35.76) that

$$\bar{y}_i = \theta_{**} + \theta_{i*} + \frac{1}{c} \sum_j \theta_{*j} + \frac{1}{c} \sum_j \theta_{ij} + \varepsilon_i, \quad (35.77)$$

where the ε_i , uncorrelated with zero mean, have variances $(\sigma^2/c^2) \sum_j n_{ij}^{-1} = \sigma^2 V_i$,

say. If we use equal weights instead of the frequency weights in (35.33), the two summations on the right of (35.77) are each equal to zero and we have simply

$$\bar{y}_i = \theta_{**} + \theta_{i*} + \varepsilon_i. \quad (35.78)$$

(35.78) is the one-way classification model (Example 35.1) with a single observation in each group, except that the error variances are not equal. If we define $z_i = \bar{y}_i/V_i^{1/2}$, we have $E(z_i) = (\theta_{**} + \theta_{i*})/V_i^{1/2}$, and the conditions of Example 35.1 are otherwise satisfied. The effect of this on the analysis is to replace S_2 defined at (35.21), which in our present application would be $\sum_i \left(\bar{y}_i - \frac{1}{r} \sum_i \bar{y}_i \right)^2$, by the same sum with each term given the coefficient V_i^{-1} , i.e.

$$S_2 = \sum_{i=1}^r \left(\sum_{j=1}^c n_{ij}^{-1}/c^2 \right)^{-1} (\bar{y}_i - \bar{y})^2$$

where

$$\bar{y} = \sum_{i=1}^r \left(\sum_{j=1}^c n_{ij}^{-1}/c^2 \right)^{-1} \bar{y}_i / \sum_{i=1}^r \left(\sum_{j=1}^c n_{ij}^{-1}/c^2 \right)^{-1}$$

is the weighted mean of the \bar{y}_i , using V_i^{-1} as weights.

We therefore have an AV table as follows:

AV for $r \times c$ cross-classification using equal weights

Variation	D.fr.	SS
Due to rows	$r-1$	$\sum_{i=1}^r V_i^{-1} (\bar{y}_i - \bar{y})^2$
Due to remainder of classification	$\frac{r(c-1)+1}{rc}$	[Obtainable as a difference]
Residual	$n-rc$	As in (35.75)
TOTAL	n	$\sum_i \sum_j \sum_p y_{ijp}^2$

(35.79)

An exactly analogous breakdown of the "classification" SS can be made for the columns-classification. We therefore have tests of row- and of column-effects in the general case. "Rows" and "Columns" SS cannot be added, because of their non-orthogonality, so that we cannot obtain the Interactions SS by differencing. However, if r or $c = 2$, a test for interactions is easily derived by this method—cf. Exercise 35.5.

35.32 The equal-weights system, whose use permitted the development of the AV (35.79), is used naturally in this context, since in effect we reduce the n observations to a set of rc means and then analyse these as though they were individual observations. There is nothing, of course, to prevent a full analysis using this equal-weights system, instead of that in 35.29, and indeed this is the customary procedure. If this is done, the results in general are different from those of 35.29.

35.33 The method used in 35.31 is due to Yates (1934), who called it the *method of weighted squares of means*. He also discusses other, more approximate, methods of analysis, as also do Snedecor (1946) and R. L. Anderson and Bancroft (1952). All these authors give numerical examples (cf. Exercise 35.7). Scheffé (1959) gives further theoretical details of the disproportional-frequencies analysis; in particular he allows arbitrary weight-systems in (35.33).

Empty cells in the two-way cross-classification

35.34 Throughout our analysis of the two-way cross-classification in 35.15–33, we made the implicit assumption that every cell in the table (35.28) contained at least one observation, i.e. $n_{ij} > 0$. In practice, it quite frequently occurs that this assumption is not fulfilled, as a result of accident, experimental failure, or other causes. We must now consider the effects which the presence of empty cells in the classification will have on the analysis of the observations. If there is at least one empty cell, the cross-classification is called *incomplete*. We have so far discussed only *complete* cross-classifications.

We clearly cannot estimate the mean μ_{ij} of an empty cell in the general case where the corresponding interaction θ_{ij} is non-zero, for we can get no information on θ_{ij} from other cells in the table. It follows from the definitions (35.31) and the linear relations (35.33) that none of the θ_{ij} , θ_{i*} , θ_{*j} or θ_{**} can be estimated in the general case if there are one or more empty cells in the cross-classification. However, even in this case we can estimate the error variance quite easily. If we denote the number of cells containing observations by $[rc]$, we obtain the more general form of (35.75):

AV for any $r \times c$ cross-classification^()*

Variation	D.fr.	SS	
Due to classification	$[rc]$	$\sum_i \sum_j n_{ij} y_{ij}^2$	
Residual	$n - [rc]$	$\sum_i \sum_j \sum_p (y_{ijp} - y_{ij})^2$	(35.80)
TOTAL	n	$\sum_i \sum_j \sum_p y_{ijp}^2$	

35.35 If the θ_{ij} in empty cells are zero, the difficulty in 35.34 disappears. Thus, if we wish to test H_3 of (35.54) (the hypothesis that *all* interactions are zero) we may proceed, as in Example 35.4, to estimate the remaining parameters, evaluate the SS due to them, and thus obtain a Residual SS. The difference between the latter and the Residual SS in (35.80) will be attributable to interactions and have $[rc] - r - c + 1$ d.fr. Similarly, if H_3 can be *postulated*, row- or column-effects can be tested as for a complete classification by the method given in Exercise 35.4. Scheffé (1959) gives further details.

(*) In (35.80) summations range over the $[rc]$ occupied cells only.

We shall not discuss the matter further here. In 37.50-6 below, we shall be giving general methods for the analysis of linear models when there are observations missing.

Hierarchical classifications

35.36 The two-way cross-classification which we have treated at length in 35.15-35 is not the only interesting generalization of the one-way classification in Example 35.1. Suppose that, within each of the k groups there considered, there is a further one-way classification of the observations. The n_1 observations in the first group are in l_1 sub-groups, with frequencies $n_{11}, n_{12}, \dots, n_{1l_1}$ where $\sum_{h=1}^{l_1} n_{1h} = n_1$; the second group similarly has l_2 sub-groups, with frequencies $n_{21}, n_{22}, \dots, n_{2l_2}$, $\sum_{h=1}^{l_2} n_{2h} = n_2$; and so on until in the k th group there are l_k sub-groups with frequencies $n_{k1}, n_{k2}, \dots, n_{kl_k}$, $\sum_{h=1}^{l_k} n_{kh} = n_k$. It will accord better with our notational conventions if we now replace the original group frequencies n_i of Example 35.1 by $n_{i.}$, to denote summation of the sub-group frequencies n_{ih} within the original groups. Thus we have

$$\sum_{h=1}^{l_i} n_{ih} = n_{i.}.$$

This is a two-way *hierarchical classification*^(*) of the observations, the *separate* sub-grouping within each of the original groups contrasting with the *common* row-grouping of every column category in a two-way cross-classification.

Example 35.5 AV in a two-way hierarchical classification

In Example 35.1 we have already defined k parameters, one for each group. In order to investigate variation in the means θ_{ih} of the l_i sub-groups within the i th group, we use only $l_i - 1$ linearly independent parameters, for we may put

$$\sum_{h=1}^{l_i} n_{ih} \theta_{ih} = 0$$

(cf. 35.17-19 for the cross-classification) so that,^(†) as at (35.34),

$$\theta_{il_l} = \frac{1}{n_{il_l}} \sum_{h=1}^{l_l-1} n_{ih} \theta_{ih}. \quad (35.81)$$

We may now generalize the linear model in Example 35.1. We write $l = \sum_{i=1}^k l_i$, and y_{ihp} for the p th observation in the h th sub-group of the i th group. We have

(*) The alternative term "nested classification" appears to be more easily taken to imply that there is an equal number of sub-groups in each original group, and we therefore do not use it, despite its appealing cosiness.

(†) These θ_{ih} are not related to the interaction parameters in the cross-classification. Interaction problems do not arise here.

$$\mathbf{y}_{(n \times 1)} = \begin{pmatrix} \bar{y}_{111} \\ \vdots \\ y_{11n_{11}} \\ y_{121} \\ \vdots \\ y_{12n_{12}} \\ \vdots \\ y_{1l_1 1} \\ \vdots \\ y_{1l_1 n_{1l_1}} \\ y_{211} \\ \vdots \\ y_{kl_k n_{kl_k}} \end{pmatrix}, \quad \boldsymbol{\theta}_{(l \times 1)} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \\ \theta_{11} \\ \vdots \\ \theta_{1, l_1-1} \\ \theta_{21} \\ \vdots \\ \theta_{2, l_2-1} \\ \vdots \\ \theta_{k1} \\ \vdots \\ \theta_{k, l_k-1} \end{pmatrix},$$

and

$$\mathbf{X}_{(n \times l)} = \begin{pmatrix} \mathbf{X}_{(n \times k)} & \mathbf{X}_{(n_1 \times (l_1-1))} & 0 \\ \mathbf{X}_{(n_2 \times (l_2-1))} & & \\ 0 & \ddots & \\ & \mathbf{X}_{(n_k \times (l_k-1))} & \end{pmatrix}. \quad (35.82)$$

The $(n \times k)$ \mathbf{X} submatrix in (35.82) is that used in Example 35.1. Each of the other submatrices \mathbf{X} is of the form

$$\mathbf{X}_{(n_i \times (l_i-1))} = \begin{pmatrix} 1 & & & & 0 \\ \vdots & & & & \\ 1 & & & & \\ 0 & 1 & & & 0 \\ \vdots & \vdots & & & \\ 0 & 1 & & & \\ 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 1 & & \\ & & & \ddots & \\ & 0 & & & 1 \\ & & & & \vdots \\ & & & & 1 \\ -\frac{n_{i1}}{n_{il_i}} \mathbf{1} & -\frac{n_{i2}}{n_{il_i}} \mathbf{1} & \dots & -\frac{n_{i, l_i-1}}{n_{il_i}} \mathbf{1} \end{pmatrix} \begin{matrix} \left. \begin{matrix} \\ \\ \\ \end{matrix} \right\} n_{i1} \text{ rows} \\ \left. \begin{matrix} \\ \\ \end{matrix} \right\} n_{i2} \text{ rows} \\ \left. \begin{matrix} \\ \end{matrix} \right\} n_{i3} \text{ rows} \\ \vdots \\ \left. \begin{matrix} \\ \\ \end{matrix} \right\} n_{i, l_i-1} \text{ rows} \\ \left. \begin{matrix} \\ \end{matrix} \right\} n_{il_i} \text{ rows} \end{matrix} \quad (35.83)$$

ANALYSIS OF VARIANCE IN THE LINEAR MODEL

which follows at once from (35.81). (35.82-3) now give

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \begin{pmatrix} n_{1.} & & 0 \\ & n_{2.} & \\ & & \ddots \\ 0 & & & n_{k.} \end{pmatrix} & 0 \\ \begin{pmatrix} 0 & & & 0 \end{pmatrix} & \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \dots & \mathbf{A}_k \end{pmatrix} \end{pmatrix}, \quad (35.84)$$

where

$$\mathbf{A}_i = \begin{pmatrix} n_{i1} + \frac{n_{i1}^2}{n_{i.}}, \frac{n_{i1}n_{i2}}{n_{i.}}, \dots, \frac{n_{i1}n_{i, l_i-1}}{n_{i.}} \\ \vdots \\ n_{i, l_i-1} + \frac{n_{i, l_i-1}^2}{n_{i.}} \end{pmatrix}$$

is of the same form as (35.37) and therefore has the inverse, of the same form as (35.44),

$$\mathbf{A}_i^{-1} = \begin{pmatrix} \frac{1}{n_{i1}} - \frac{1}{n_{i.}}, -\frac{1}{n_{i.}}, -\frac{1}{n_{i.}}, \dots, -\frac{1}{n_{i.}} \\ \vdots \\ \frac{1}{n_{i2}} - \frac{1}{n_{i.}}, -\frac{1}{n_{i.}}, \dots, -\frac{1}{n_{i.}} \\ \vdots \\ \frac{1}{n_{i, l_i-1}} - \frac{1}{n_{i.}}, -\frac{1}{n_{i.}} \end{pmatrix}. \quad (35.85)$$

Hence, from (35.84-5), the LS estimators are

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} y_{1..} \\ y_{2..} \\ \vdots \\ y_{k..} \\ y_{11.} - y_{1..} \\ y_{12.} - y_{1..} \\ \vdots \\ y_{1, l_1-1.} - y_{1..} \\ y_{21.} - y_{2..} \\ \vdots \\ y_{k, l_k-1.} - y_{k..} \end{pmatrix}, \quad (35.86)$$

and thus the SS due to the fitted model is

$$\hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}} = \sum_{i=1}^k n_{i.}y_{i..}^2 + \sum_{i=1}^k \sum_{h=1}^{l_i} n_{ih} (y_{ih.} - y_{i..})^2. \quad (35.87)$$

The first term on the right of (35.87) is precisely S , defined at (35.16) in Example

35.1. What we have now done is to partition off a further SS, the second term on the right of (35.87). Since S_1 was the SS due to the k original group parameters $\theta_1, \theta_2, \dots, \theta_k$, the second SS is that attributable to the $\sum_{i=1}^k (l_i - 1) = l - k$ linearly independent sub-group parameters now introduced. We may summarize our result in the table:

AV for a two-way hierarchical classification

Variation	D.fr.	SS
Due to groups	k	$\sum_i n_i. y_{i..}^2$
Between sub-group within groups	$l - k$	$\sum_i \sum_h n_{ih} (y_{ih.} - y_{i..})^2$
Residual	$n - l$	$\sum_i \sum_h \sum_p (y_{ihp} - y_{ih.})^2$
TOTAL	n	$\mathbf{y}' \mathbf{y}$

The Residual SS in (35.88) is obtained by subtraction. The first row of the table may be split into a "Between groups" and "General mean" components as in Example 35.1.

The ratio of the "Between sub-groups within groups" MS to the Residual MS is, from our general theory, a non-central F -variable with d.fr. $(l - k, n - l)$ and non-central parameter $\lambda = \sum_i \sum_h n_{ih} \theta_{ih}^2$, which is zero when all sub-group means within each group are equal, giving a central F -test for this hypothesis.

35.37 The hierarchical process can clearly be carried further, with sub-sub-groups and even sub-sub-sub-groups. These would be termed three-way and four-way hierarchical classifications, and are relatively rare in practice. It should be obvious to the reader that there is no need to go through again the rather tedious algebra of LS theory to obtain the results we need here; the work of Example 35.5 essentially split the SS within each of the k groups into two components, with $(l_i - 1)$ and $(n_i - l_i)$ d.fr. respectively, and summed corresponding components over all groups to obtain the $(l - k)$ and $(n - l)$ d.fr. in the table (35.88). The same splitting-off process can now be carried out within each sub-group, and so on. The reader is asked to verify the three-way AV in Exercise 35.3. Scheffé (1959) gives theoretical details for the three-way case.

Multi-way classifications

35.38 We have just outlined the treatment of multi-way hierarchical classifications, and this leads us to a consideration of multi-way classifications in general. We first note that, as soon as we consider three-way classifications, there is the possibility of "mixed" classifications which are partly hierarchical and partly cross-

classifications. These arise when a two-way hierarchical classification forms one (say, the row-) classification of a $r \times c$ cross-classification. In the notation of Example 35.5, we here have $r = l$. The AV is carried out in two stages. First, the cross-classification is analysed by the methods already discussed, and the Total SS is resolved into the usual five components, which we represent concisely by their d.fr. in the following table:

SS due to	D.fr.	
General mean	1	
Rows	$l-1$	$k-1$
Columns	$c-1$	$l-k$
Interactions	$(l-1)(c-1)$	$(k-1)(c-1)$
Residual	$n-lc$	$(l-k)(c-1)$
TOTAL	n	

At the second stage, each of the SS involving the hierarchical (row) classification is subdivided into two parts as indicated on the right. The first of these subdivisions is a direct application of Example 35.5 (it being remembered that the general mean component has here already been removed from the first line of (35.88) by the cross-classification analysis), but the simplest way of achieving both subdivisions is to merge all sub-groups within the groups of the hierarchical classification and recalculate the SS for Rows and Interactions using the merged data—these are the required component SS, with $(k-1)$ and $(k-1)(c-1)$ d.fr. respectively. The sub-groups SS are then obtained as differences if the analysis is orthogonal.

Scheffé (1959) gives theoretical details for the case where there is the same number of sub-groups in each group of the hierarchical classification and the same number of observations in each cell of the $l \times c$ table.

35.39 Suppose now that, instead of embedding a two-way hierarchical classification within a two-way cross-classification as in 35.38, we carry out a new one-way classification within each cell of a cross-classification. If the *same* one-way classification is carried out in each cell, we clearly arrive at a three-way cross-classification. All the problems of formulating the linear model, discussed in 35.15–19 for the two-way case, now arise afresh, and some generalization of our concepts is required, as we shall now see.

The three-way cross-classification

35.40 Following the nomenclature already used in the treatment of three-way tables of categorized data in 33.58, we now consider a sample of n observations classified into a $r \times c \times l$ table with r rows, c columns and l “layers,” with frequencies n_{ijk} , where $i = 1, 2, \dots, r$; $j = 1, 2, \dots, c$; $k = 1, 2, \dots, l$. The p th observation in

the (i, j, k) th cell is y_{ijkp} , $p = 1, 2, \dots, n_{ijk}$. As in 35.15, we follow the notational rule that a dot replacing a suffix indicates *summation* if the suffix is to n , and *averaging* if the suffix is to y , and we replace $n_{...}$ by n .

We can clearly ask the questions of 35.15 about all three classifications in this more general situation, so that we are now interested in the general mean, in row-effects, column-effects and layer-effects (the three sets of *main effects*), and in the interactions between any pair of the row-, column- and layer-classifications. However, there is a new feature introduced by the additional classification, for we may now wish to know whether the interactions between, say, rows and columns themselves depend upon the layer-classification. We are here concerned with a higher-order interaction, which we must proceed to define.

35.41 We write the linear model

$$y_{ijkp} = \mu_{ijk} + \varepsilon_{ijkp} = \theta_{***} + \theta_{i**} + \theta_{*j*} + \theta_{**k} + \theta_{ij*} + \theta_{i*k} + \theta_{*jk} + \theta_{ijk} + \varepsilon_{ijkp}, \quad (35.89)$$

the generalization of (35.32). μ_{ijk} , the mean of the observations in the (i, j, k) th cell, is made up of eight components: the general mean θ_{***} ; the row-effects θ_{i**} , column-effects θ_{*j*} , and layer-effects θ_{**k} ; the row-column interactions θ_{ij*} , the row-layer interactions θ_{i*k} , and the column-layer interactions θ_{*jk} . All these are defined exactly as in 35.17-18 for the two-way cross-classification, with an extra asterisk in every suffix. The last set of interaction parameters on the right of (35.89), the θ_{ijk} , are defined by extending the argument of 35.18. If the deviation of μ_{ijk} from the general mean $\theta_{***} \equiv \mu_{***}$ were exactly equal to the sum of the three corresponding main effects and the three corresponding interactions already defined, we should have

$$\begin{aligned} \mu_{ijk} - \mu_{***} &= \theta_{i**} + \theta_{*j*} + \theta_{**k} + \theta_{ij*} + \theta_{i*k} + \theta_{*jk} \\ &= (\mu_{i**} - \mu_{***}) + (\mu_{*j*} - \mu_{***}) + (\mu_{**k} - \mu_{***}) \\ &\quad + (\mu_{ij*} - \mu_{i**} - \mu_{*j*} + \mu_{***}) + (\mu_{i*k} - \mu_{i**} - \mu_{**k} + \mu_{***}) \\ &\quad + (\mu_{*jk} - \mu_{*j*} - \mu_{**k} + \mu_{***}), \end{aligned}$$

or $\theta_{ijk} = 0$ where (compare the definition of θ_{ij} in (35.31))

$$\theta_{ijk} = \mu_{ijk} - (\mu_{ij*} + \mu_{i*k} + \mu_{*jk}) + (\mu_{i**} + \mu_{*j*} + \mu_{**k}) - \mu_{***}. \quad (35.90)$$

We call the θ_{ijk} as defined by (35.90) the *second-order interactions* between rows, columns and layers. The interactions in a two-way cross-classification are now retrospectively re-defined as *first-order interactions*. For complete terminological regularity, we may also refer to the main effects themselves as *zero-order interactions*.

If we write (35.90) in the form

$$\theta_{ijk} = \{\mu_{ijk} - (\mu_{i*k} + \mu_{*jk}) + \mu_{**k}\} - \{\mu_{ij*} - (\mu_{i**} + \mu_{*j*}) + \mu_{***}\}, \quad (35.91)$$

and refer to the definition of θ_{ij} at (35.31), we see that θ_{ijk} is in fact the difference between a row-column (first-order) interaction in the k th layer and the same interaction in all layers combined. Because θ_{ijk} is symmetrically defined as the difference between a row-layer and layers, (35.91) may be written equivalently as the difference between a row-layer interaction in the j th column and in all columns combined, or as the difference between a column-layer interaction in the i th row and in all rows combined. Thus the word "interaction" is as apposite for θ_{ijk} here as it was for θ_{ij} in 35.18.

ANALYSIS OF VARIANCE IN THE LINEAR MODEL

35.42 Analogously with 35.17, we can only have rc linearly independent parameters in our three-way cross-classification, one for each cell in the table. The model (35.89), however, contains

$$1 + r + c + l + rc + rl + cl + rcl$$

parameters, and the surplus must be dropped to avoid singularity in the model. We therefore drop the last of each set of main-effect parameters, and reduce the sets of first-order interaction parameters to $(r-1)(c-1)$, etc., just as in 35.17. This gives us

$$1 + (r-1) + (c-1) + (l-1) + (r-1)(c-1) + (r-1)(l-1) + (c-1)(l-1)$$

parameters excluding the second-order interactions. The reader may verify by addition that $(r-1)(c-1)(l-1)$ of these bring the total number of parameters to the required number rc —this is otherwise obvious from the fact that this is the number of μ_{ijk} which can be independently determined when the other parameters are known.

35.43 Nothing but the heavy algebra now prevents our following through in detail an analysis parallel to that already carried out in the two-way case. There is no numerical difficulty in any given case about fitting the linear model (35.89), estimating its rc parameters and carrying out the AV. Even in the two-way case, however, worthwhile simplifications in the algebra only occurred when the frequency in each cell was proportional to the product of marginal totals as required by (35.43). Similar orthogonality conditions now require that each cell frequency should be proportional to the product of all the corresponding marginal frequencies (cf. Seber (1964a)). This proportionality condition, in practice, is satisfied with equal frequencies in all the cells, i.e. in the balanced case.

The general principles of the analysis are then very simply set out. We saw in 35.39 that we may regard a three-way ($r \times c \times l$) cross-classification as having been generated by imposing a new (l -fold) classification upon every cell of an existing ($r \times c$) cross-classification. It follows that the AV can be carried out in two stages, exactly as for the "mixed" classification of 35.38. First, we consider the ($r \times c$) cross-classification as a one-way classification with rc cells, and carry out the AV of the $rc \times l$ two-way cross-classification in which our observations are then displayed. We obtain the schematic AV table from (35.65):

SS due to	D.fr.	
General mean	1	
$(r \times c)$ cross-classification	$rc - 1$	$\left\{ \begin{array}{l} \text{rows} \quad r - 1 \\ \text{columns} \quad c - 1 \\ \text{row-column} \\ \text{interactions} \end{array} \right. \quad (r-1)(c-1)$
Layer classification	$l - 1$	
Interactions of $(r \times c)$ cross-classification with l layers	$(rc - 1)(l - 1)$	$\left\{ \begin{array}{l} \text{row-layer} \\ \text{interactions} \quad (r-1)(l-1) \\ \text{column-layer} \\ \text{interactions} \quad (c-1)(l-1) \\ \text{row-column-layer} \\ \text{second-order interactions} \end{array} \right. \quad (r-1)(c-1)(l-1)$
Residual	$n - rcl$	

(35.92)

At the second stage, each of the two SS involving the $(r \times c)$ cross-classification is subdivided into three parts as shown on the right. The simplest method (as in 35.38) of doing this is first to merge all columns within rows and recalculate the two SS, which then are the required components with $(r-1)$ and $(r-1)(l-1)$ d.fr.; if the merging operation is now separately applied to all rows within columns, and the two SS again recalculated, we obtain the required components with $(c-1)$ and $(c-1)(l-1)$ d.fr. The two remaining SS (with $(r-1)(c-1)$ and $(r-1)(c-1)(l-1)$ d.fr.) are now obtainable as differences since the analysis is orthogonal.

The computation of the AV in the general disproportional frequencies case remains formidable, even with electronic computers. Pearce (1963) reviews the situation generally. Freeman and Jeffers (1962) give a method for the non-orthogonal three-way cross-classification; Stevens (1953) used an iterative method for this case. Bradu (1965) solves the problem for the simple case where all interactions of all orders are assumed zero—see also Rees (1966).

Gabriel (1963) gives an expository review of the theory for analysing cell means, whose variances are inversely proportional to cell sample sizes, with special reference to the case when y is a 0-1 variable, and the cell means become proportions. An approximate method of analysing cell means is given in Exercise 37.7.

Example 35.6 *Balanced three-way cross-classification*

In the case where all cell-frequencies are equal to $m > 1$, it is at once obvious that we may treat the $(r \times c \times l)$ cross-classification as a $(r \times c) \times l$ or a $(r \times l) \times c$ or a $(c \times l) \times r$ at the first stage of calculating the AV in (35.92). It will be seen from this symmetry that each of the three sets of main effects and each of the three sets of first-order interactions will have its SS calculated exactly as in a two-way cross-classification table with the third factor of classification merged. The Residual SS is clearly also unchanged in form. In our present three-way notation, we therefore obtain the following expressions from (35.63-4) and Exercise 35.1 for the SS corresponding to the components in (35.92). The suffix to S now indicates its d.fr.

General mean:

$$S_1 = (\sum_i \sum_j \sum_k \sum_p y_{ijkp})^2 / (rclm),$$

Row-effects:

$$S_{(r-1)} = \sum_i (\sum_j \sum_k \sum_p y_{ijkp})^2 / (clm) - S_1,$$

Column-effects:

$$S_{(c-1)} = \sum_j (\sum_i \sum_k \sum_p y_{ijkp})^2 / (rlm) - S_1,$$

Layer-effects:

$$S_{(l-1)} = \sum_k (\sum_i \sum_j \sum_p y_{ijkp})^2 / (rcm) - S_1,$$

Row-column interactions:

$$S_{(r-1)(c-1)} = \sum_i \sum_j (\sum_k \sum_p y_{ijkp})^2 / (lm) - S_{(r-1)} - S_{(c-1)} - S_1,$$

Row-layer interactions:

$$S_{(r-1)(l-1)} = \sum_i \sum_k (\sum_j \sum_p y_{ijkp})^2 / (cm) - S_{(r-1)} - S_{(l-1)} - S_1,$$

(35.93)

ANALYSIS OF VARIANCE IN THE LINEAR MODEL

$$\left. \begin{array}{l} \text{Column-layer interactions:} \\ S_{(c-1)(l-1)} = \sum_j \sum_k (\sum_i y_{ijkp})^2 / (rm) - S_{(c-1)} - S_{(l-1)} - S_1, \\ \text{Residual:} \\ S_{rcl(m-1)} = \sum_i \sum_j \sum_k \sum_p y_{ijkp}^2 - \sum_i \sum_j \sum_k (\sum_p y_{ijkp})^2 / m. \end{array} \right\} (35.93) \text{ cont.}$$

Since the eight components in (35.93) together with the SS for the second-order interaction must add to $\sum_i \sum_j \sum_k \sum_p y_{ijkp}^2$, as always, we obtain by subtraction the SS:

Second-order interaction:

$$S_{(r-1)(c-1)(l-1)} = \sum_i \sum_j \sum_k (\sum_p y_{ijkp})^2 / m - S_{(r-1)(c-1)} - S_{(r-1)(l-1)} - S_{(c-1)(l-1)} - S_{(r-1)} - S_{(c-1)} - S_{(l-1)} - S_1. \quad (35.94)$$

We finally assemble the results of (35.92-4) into the AV table below.

Variation due to	D.fr.	SS defined by (35.93-4)
Row-effects	$r-1$	$S_{(r-1)}$
Column-effects	$c-1$	$S_{(c-1)}$
Layer-effects	$l-1$	$S_{(l-1)}$
Row-column interactions	$(r-1)(c-1)$	$S_{(r-1)(c-1)}$
Row-layer interactions	$(r-1)(l-1)$	$S_{(r-1)(l-1)}$
Column-layer interactions	$(c-1)(l-1)$	$S_{(c-1)(l-1)}$
Row-column-layer interactions	$(r-1)(c-1)(l-1)$	$S_{(r-1)(c-1)(l-1)}$
General mean	$rcl-1$ 1	S_1
Classification	rcl	
Residual	$rcl(m-1)$	$S_{rcl(m-1)}$
TOTAL	$rclm=n$	$y'y$

(35.95)

Any of the eight rows of (35.95) forming part of the "Classification" SS may be tested against the Residual SS, just as previously, by the ratio of its SS/d.fr. to the Residual SS/d.fr. Each ratio has a non-central F distribution, becoming central if the hypothesis tested holds.

Multi-way cross-classifications

35.44 The reader should now be able to see how the three-way cross-classification analysis can be further generalized to four- and more-way classifications by repeated application of the argument we used to obtain the three- from the two-way analysis. The formal symmetry of (35.95) in the balanced case, and also of (35.93-4), invite more direct generalization to higher-order classifications. We should notice particularly

the uniform correspondence of the d.fr. attached to an SS and the number of linearly independent parameters which it represents—this is easily seen as a consequence of geometrical arguments of the type referred to in 35.13 above, which are discussed in detail by Scheffé (1959).

Any such further generalization involves the definition of third- and higher-order interactions if these are required in the model, but these interactions tend to be so remote and difficult to interpret that they are frequently ignored in the subsequent analysis, their SS and d.fr. being merged with the Residual.

The combination of AV tests

35.45 Suppose first quite generally that k distinct hypotheses were to be tested, and that their respective test statistics were all independently distributed. To obtain a combined test of the k hypotheses, we could use the result of Exercise 16.4 in the manner of Exercise 30.9, Vol. 2. Applying the probability-integral transformation to each test statistic, and directing it so that critical values of each test statistic correspond to small values of its transform P_i , we then have $-2 \sum_{i=1}^k \log P_i = P$ as a χ^2 variable with $2k$ d.fr., large values of P being critical. Whatever the sizes of the constituent tests, we use a size- α test on P . If the tests are not independent, however, we encounter exactly the difficulties mentioned in the context of tests of fit in 30.36, and this combined test is not useful since its distribution requires knowledge of the joint distribution of the test statistics.

Another simple general approach to the combination of independent tests arises from the observation that if the i th test has size α_i , the probability of rejecting at least one of the hypotheses tested when all are true is simply $1 - \prod_{i=1}^k (1 - \alpha_i)$, which reduces when all $\alpha_i = \alpha$ to

$$P_k(\alpha) = 1 - (1 - \alpha)^k, \quad (35.96)$$

approximately equal to $k\alpha$ when α is small, as it normally is in practice.

35.46 Now if all the k tests in an AV table were independent, we could use (35.96) to fix the overall size in testing the set of variance-ratios as a whole, so that if there are four tests to be made at size α , and we required overall size to be 0.05, we should solve

$$0.05 = 1 - (1 - \alpha)^4$$

for α or, approximately, put $\alpha = 0.05/4$. However, the tests in the AV tables which we have considered are never independent tests, for although the various SS in a table may be independent of each other, all the tests we have derived use the Residual SS as denominator of the test statistic, and the various tests must therefore be statistically dependent, since, e.g., a Residual SS which is (by chance) large will depress the values of all the test statistics simultaneously.

35.47 Fortunately, however, (35.96) still holds as an approximation, as Hartley (1955) showed. Suppose that $(k+1)$ independent mean squares s^2, s_1^2, \dots, s_k^2 are

ANALYSIS OF VARIANCE IN THE LINEAR MODEL

observed, with respective d.f. ν, ν_1, \dots, ν_k . We write G, G_1, \dots, G_k for their distribution functions and $g = G'$ for the f.f. of s^2 . Let the k values F_i be defined as the solutions for a fixed α of

$$\text{Prob} \{s_i^2/s^2 \leq F_i\} = 1 - \alpha,$$

so that F_i is the $100(1-\alpha_i)$ per cent quantile of the distribution of the ratio s_i^2/s^2 . The probability that none of the ratios s_i^2/s^2 exceeds its F_i is

$$P(1) = \text{Prob} \left\{ \frac{s_1^2}{s^2} \leq F_1, \dots, \frac{s_k^2}{s^2} \leq F_k \right\} = \int_0^\infty \left\{ \prod_{i=1}^k G_i(xF_i) \right\} g(x) dx. \quad (35.97)$$

We have denoted (35.97) by $P(1)$ because it is the value at $\theta = 1$ of the function

$$P(\theta) = \int_0^\infty \prod_{i=1}^k [(1-\alpha) + \theta \{G_i(xF_i) - (1-\alpha)\}] g(x) dx, \quad (35.98)$$

and we see at once that $P(0) = (1-\alpha)^k$. In order to expand $P(1)$ in a Taylor series about zero, we investigate its derivatives. We find

$$P'(\theta) = \sum_{i=1}^k \int_0^\infty \{G_i(xF_i) - (1-\alpha)\} \prod_{\substack{j=1 \\ j \neq i}}^k [(1-\alpha) + \theta \{G_j(xF_j) - (1-\alpha)\}] g(x) dx,$$

so that

$$P'(0) = (1-\alpha)^{k-1} \sum_i \int_0^\infty \{G_i(xF_i) - (1-\alpha)\} g(x) dx = 0$$

since

$$\int_0^\infty G_i(xF_i) g(x) dx = 1 - \alpha. \quad (35.99)$$

Thus the Taylor expansion is

$$\begin{aligned} P(1) - (1-\alpha)^k &= \frac{1}{2} P''(\bar{\theta}), \quad 0 \leq \bar{\theta} \leq 1, \\ &= \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^k \int_0^\infty \{G_i(xF_i) - (1-\alpha)\} \{G_j(xF_j) - (1-\alpha)\} \\ &\quad \times \prod_{\substack{l=1 \\ l \neq i,j}}^k [(1-\alpha) + \bar{\theta} \{G_l(xF_l) - (1-\alpha)\}] g(x) dx. \end{aligned}$$

Every term in square brackets lies between 0 and 1, since $\alpha, \bar{\theta}$ and G do so, and if we assume all these terms equal to 1 we therefore obtain the inequality

$$P(1) - (1-\alpha)^k \leq \frac{1}{2} \sum_{i \neq j} \int_0^\infty \{G_i(xF_i) - (1-\alpha)\} \{G_j(xF_j) - (1-\alpha)\} g(x) dx,$$

from which the Cauchy-Schwarz inequality gives the blunter inequality

$$\begin{aligned} |P(1) - (1-\alpha)^k| &\leq \frac{1}{2} \sum_{i \neq j} \left[\int_0^\infty \{G_i(xF_i) - (1-\alpha)\}^2 g(x) dx \int_0^\infty \{G_j(xF_j) - (1-\alpha)\}^2 g(x) dx \right]^{\frac{1}{2}}, \end{aligned}$$

and, even further,

$$|P(1) - (1-\alpha)^k| \leq \frac{1}{2} k(k-1) \max_i \int_0^\infty \{G_i(xF_i) - (1-\alpha)\}^2 g(x) dx,$$

and because of (35.99) we may write this

$$|P(1) - (1-\alpha)^k| \leq \frac{1}{2} k(k-1) \max_i [\text{var} \{G_i(s^2 F_i)\}]. \quad (35.100)$$

35.48 Hartley (1955) has computed the upper bound given by (35.100) for a range of values of k , ν and ν_i when $(1-\alpha)$ is near 0.05. It decreases to zero as ν increases to infinity, but increases with k and ν_i . For the rather unfavourable case $k = 10$, $\nu = 30$, $\max_i \nu_i = 10$, the upper bound is 0.0138; if ν is then increased to 60, the bound drops to 0.0050. For $k = 5$, $\nu = 60$, $\max_i \nu_i = 5$, the bound is 0.0014. Since

the actual difference will generally be appreciably less than the upper bound (three separate "throwing-away" processes produced the bound), we may conclude that the approximation of (35.97) by $(1-\alpha)^k$ is quite satisfactory over a wide range, especially when ν (the d.f. for the Residual SS) is large.

We therefore see that the set of k variance-ratios in an AV table is tested with size $P_k(\alpha) = 1 - P(1)$, related approximately by (35.96) to the nominal size α at which each variance-ratio in the table is tested. Alternatively, if we rewrite (35.96) in the form

$$1 - \alpha = \{1 - P_k(\alpha)\}^{1/k}$$

and substitute α for $P_k(\alpha)$ and $\beta(k)$ for α , we have

$$\beta(k) = 1 - (1 - \alpha)^{1/k}, \quad (35.101)$$

so that a test of size α in the set of k variance-ratios is achieved if each of them is separately tested with size $\beta(k)$ defined from α by (35.101).

Earlier work on this problem was done by Hartley (1938) and Finney (1941). Cochran (1941) and Darling (1952) considered the ratio of the largest mean square to the sum of all mean squares when all ν_i are equal.

35.49 The result of 35.47-8 states that separate tests of size (35.101) on the k variance-ratios with common Residual SS give an overall test of approximate size α . However, the fact that the denominator SS is common to all k tests suggests that it may be inefficient to make the tests separately in this way, since the result of any one of the tests gives relevant information for the others. A step-by-step procedure which utilizes this fact was suggested by Hartley (1955).

Define H_i as the probability-integral transformation of s_i^2/s^2 to the uniform distribution on $(0, 1)$ (cf. 1.27), and order the H_i so that $H_{(i)}$ is the i th smallest of them. Since $\text{Prob}\{H_i \geq 1 - \delta\} = \delta$, a test of size δ on H_i is obtained by comparing it with $1 - \delta$.

The first step is to test $H_{(k)}$, the largest of the H_i , with size $1 - (1 - \alpha)^{1/k} = \beta(k)$ as at (35.101). If $H_{(k)} < 1 - \beta(k)$, the hypothesis of homogeneity of all the observations is accepted outright; if $H_{(k)} \geq 1 - \beta(k)$, we reject the corresponding hypothesis of "no effect" in the AV table, and proceed to test $H_{(k-1)}$ at size $\beta(k-1)$. If $H_{(k-1)} < 1 - \beta(k-1)$, we accept all the $(k-1)$ remaining hypotheses of "no effect" in the table; if $H_{(k-1)} \geq 1 - \beta(k-1)$, we reject the corresponding hypothesis and proceed to test $H_{(k-2)}$ at size $\beta(k-2)$, and so on until some $H_{(i)} < 1 - \beta(i)$ (when the i remaining hypotheses are accepted) or $H_{(1)} \geq 1 - \beta(1)$, by which time all k "no effect" hypotheses will have been rejected.

35.50 This step-by-step test is easily shown to have size α . Suppose that of the k variance-ratios, c correspond to "zero effects" and $(k-c)$ to non-zero effects, and that

of the c transformed values H_i corresponding to the former among the k H_i , $H_{(l)}$ is the largest. If any of the "zero-effects" hypotheses are rejected by the test procedure, that corresponding to $H_{(l)}$ must be, since it is reached first in the step-by-step process. The probability that any "zero-effect" hypothesis is rejected is therefore

$$P = \text{Prob} \{H_{(i)} \geq 1 - \beta(i), \quad i = l, l+1, \dots, k\} \\ \leq \text{Prob} \{H_{(l)} \geq 1 - \beta(l)\}.$$

Since $l \geq c$, and $\beta(i)$ is a decreasing function of i , we have further

$$P \leq \text{Prob} \{H_{(l)} \geq \{1 - \beta(c)\} = \alpha, \quad (35.102)$$

since if $H_{(l)}$, the largest of a set of c variance-ratios, is tested with size $\beta(c)$ we obtain an overall test of size α by (35.101). (35.102) shows that the step-by-step test never has size exceeding α . If $c = k$, $l = k$ also and (35.102) becomes an equality.

Hartley (1955) gives some consideration to the power of the test when all mean squares have equal d.f.

Multiple comparisons

35.51 Each of the variance-ratio tests in an AV table tests a hypothesis concerning a set of parameters, e.g. the row-effects or the interactions between rows and columns. For practical purposes, however, it is often not enough to know, for example, that the row-effects θ_{i*} are different—we need to know which of the θ_{i*} are to be regarded as greater than the others, or more generally whether the θ_{i*} may be said to fall into distinct groups.

Now the LS estimators $\hat{\theta}_{i*}$ are, of course, the MV unbiased linear estimators of their corresponding parameters, and provide us with estimators of any of the differences $\theta_{i*} - \theta_{j*}$, but we are usually unable to nominate the differences of interest in advance, and we therefore are faced with the problem of carrying out a number of non-independent tests on the differences. The discussion of 35.45–6 applies here with obvious changes, although the problem we are now concerned with is a more detailed one. Whereas in 35.47–50 we were dealing with combining tests on sets of parameters, we are now interested in closer examination of a particular set, say the row-effects.

In a sense, this problem of *multiple comparisons*, as it is called, is a more complex version of the problem of outlying observations, discussed in 32.23–8. Instead of being concerned about a location-shift in one or more observations, we are now more generally asking whether the observations (here the $\hat{\theta}_{i*}$) have expected values (the θ_{i*}) which fall into distinct groups. Tukey (1953) reviews the subject.

The LSD test

35.52 For the sake of definiteness, our discussion will refer to a one-way classification as in Example 35.1, although there is no essential difference if we consider any set of effects in an AV table. In Example 35.1, the observed group means y_i are the LS estimators of the k group parameters θ_i (each of which includes the general mean as a common element). If the F -test at (35.22) rejects the hypothesis (35.20) that the θ_i are all equal, we are faced with the need to decide which subsets of the θ_i may be regarded as homogeneous, and which not.

The simplest test procedure is the oldest ("Student," 1908), namely to carry out

an ordinary two-sample "Student's" t -test on every one of the $\frac{1}{2}k(k-1)$ possible pairs of $y_i, y_j, i \neq j$. If each of these tests has size α , we can say little about the size of the overall combined test, since they are non-independent—no simple combination formula like (35.96) is available.

This combined test amounts to calculating an estimated standard error of the difference between two means (using the Residual SS) and comparing the observed difference with it, using the appropriate (Residual SS) number of d.f. in the "Student's" distribution. If the number of observations in each group (n_i) is the same (equal to N , say) only a single standard error estimate is needed, of form $(2s^2/N)^{\frac{1}{2}}$, where s^2 is the unbiased estimator of the error variance σ^2 . We thus set up a *Least Significant Difference* (the appropriate multiple of the standard error for a two-sided test of size α) and compare each of the $\frac{1}{2}k(k-1)$ observed differences with it. In consequence, this is sometimes called the *LSD test*. One cannot say much more in general about the LSD test than that if the true group means do not differ, a proportion α of all pairs adjudged heterogeneous by the test will be wrongly so judged.

A simple modification of the LSD test, proposed by Fisher (1935), is to reduce the size of each component test from α to $\alpha / \binom{k}{2}$. This has the effect of reducing the expected number of pairs erroneously adjudged heterogeneous (when all group means are truly equal) from $\frac{1}{2}k(k-1)\alpha$ to simply α .

When using either the original or modified LSD test, it must be remembered that the expected error rates just referred to are unconditional ones, taking no account of the fact that the test is made after (and because) the overall F -test rejected the hypothesis of homogeneity.

Step-by-step and simultaneous test procedures

35.53 Like the outlier problem of 32.23-5, which it resembles, the multiple comparisons problem was often discussed in terms of sample range criteria.

The simplest of these is Tukey's (1951, 1952) studentized range test. The k group means, which we shall now write $\bar{x}_i, i = 1, 2, \dots, k$, are (on the hypothesis of homogeneity) a random sample of size k from a normal distribution with variance σ^2/N , independently estimated by s^2/N , where N is the number of observations in each group as before. Suppose the group means to be ordered, and denote them by $\bar{x}_{(1)} \leq \bar{x}_{(2)} \leq \dots \leq \bar{x}_{(k-1)} \leq \bar{x}_{(k)}$. To any pair $\bar{x}_{(i)}$ and $\bar{x}_{(j)}, i < j$, there corresponds a difference $(\bar{x}_{(j)} - \bar{x}_{(i)})$ which is the range of a subset of $(j-i+1)$ adjacent ordered group means. This subset is adjudged heterogeneous (and the extreme group means $\bar{x}_{(i)}$ and $\bar{x}_{(j)}$ therefore regarded as from different populations) if $(\bar{x}_{(j)} - \bar{x}_{(i)}) / (s/N^{\frac{1}{2}})$ exceeds the 100(1- α) per cent point, q_α , of the studentized range (cf. 32.25) of k observations. Since no subset range can do this unless the range $(\bar{x}_{(k)} - \bar{x}_{(1)})$ does, the procedure clearly has overall size α .

In practice, $\bar{x}_{(k)}$ is successively compared with $\bar{x}_{(1)}, \bar{x}_{(2)}$, and so on until a "homogeneous" verdict is reached. If $(\bar{x}_{(k)} - \bar{x}_{(1)})$ is homogeneous, the test ends there; otherwise all means $\bar{x}_{(1)}, \bar{x}_{(2)}, \dots$ found to be heterogeneous from $\bar{x}_{(k)}$ are tested against $\bar{x}_{(k-1)}$, again in succession starting from $\bar{x}_{(1)}$. All means then found to be heterogeneous

from $\bar{x}_{(k-1)}$ are tested against $\bar{x}_{(k-2)}$, and so on until no further "heterogeneous" verdict can be reached. Such procedures, in which the decision on each subset depends on previous decisions concerning a larger subset, are called *step-by-step* or *stepwise* procedures.

35.54 A different *simultaneous test procedure* employing a sum-of-squares, rather than a range, technique is suggested by Gabriel (1964). The "between-groups" SS (35.21) is calculated for *every one* of the $(2^k - k - 1)$ subsets of two or more of the k groups, and tested against the *fixed* critical value obtained from the variance-ratio distribution with $(k-1, n-k)$ d.fr. The sample sizes n_i need not now be equal. This procedure leads to transitive judgements in the sense that no subset can be adjudged heterogeneous when a larger subset containing it is not; i.e. no subset can be adjudged homogeneous unless all smaller subsets which it contains are also homogeneous (cf. Exercise 35.16). Clearly, this procedure contains the ordinary variance-ratio test as a component, when the subset is actually the whole set of k groups. This implies that the test has overall size α .

When all the n_i are equal, Tukey's test in 35.53 can be modified to be a "simultaneous test procedure" of the same type as Gabriel's if every subset of, instead of merely every subset of *adjacent*, group means is tested by the range criterion. (This test has the additional property that if any set of more than two groups is adjudged heterogeneous, at least one subset of it would be.)

Both the Gabriel and the Tukey-Gabriel methods discussed in this section have the property that the probability of erroneously judging a subset to be heterogeneous decreases with the size of the subset. For $k = 8$ and 40 d.fr. for the Residual SS, this phenomenon is more marked for the former method—Gabriel (1964) gives tables. The Tukey-Gabriel method is much simpler computationally, especially for large k , but is only available when all n_i are equal.

35.55 Instead of using the fixed critical value of the k -observations studentized range, as in 35.53, we can test $(x_{(j)} - x_{(i)})/(s/N^{1/2})$ against the studentized range of $(j-i+1)$ observations, as suggested earlier by Newman (1939) and Keuls (1952). A new point now arises, for a set of q adjacent (ordered) group means may be declared "homogeneous" while a subset of p adjacent group means, contained within that set of q , is "heterogeneous" by this criterion. The Newman-Keuls step-by-step procedure adjudges a pair of group means heterogeneous only if *every* subset of adjacent group means containing that pair is heterogeneous by the studentized range test just defined, which takes account of the number in the subset.

The computational procedure is just as in the last paragraph of 35.53, except that the critical value in the studentized range test now varies in the component tests, instead of being fixed as previously. Once again, the overall size is at once seen to be α , since $(x_{(k)} - x_{(1)})$ must first be adjudged heterogeneous if any other difference is to be.

D. B. Duncan (1952, 1955, 1957) proposes what is essentially a modification of the Newman-Keuls procedure in which each difference $(x_{(j)} - x_{(i)})$ is tested against the $100(1 - \alpha_{j-i+1})$ per cent point of the studentized range of $(j-i+1)$ observations, where

α_{j-i+1} now depends on $(j-i)$ through the relation

$$\alpha_{j-i+1} = 1 - (1 - \alpha_2)^{j-i}, \quad (35.103)$$

the argument stemming from (35.96) and the consideration that a test of $(j-i+1)$ observations is equivalent to $(j-i)$ separate tests on pairs of observations.

Thus the probabilities of error are redistributed among the components of the test procedure, falling as the group means compared are closer together in the ordering (cf. 35.54). D. B. Duncan (1955) provides tables for overall test size $\alpha = 0.05$ and 0.01 .

35.56 D. B. Duncan (1955), Tukey (1953) and Scheffé (1959) discuss some other multiple comparison procedures, and Hartley (1957) makes some power comparisons, while Hartley (1955) briefly considers the power of the Newman-Keuls method. Gabriel (1964) compares the Newman-Keuls and Duncan step-by-step methods with the two simultaneous test procedures discussed in 35.54, and proves that for given overall test size and any step-by-step method, the probability of erroneously judging any subset heterogeneous cannot be less than for the simultaneous test procedure based on the same statistic.

Simultaneous confidence intervals for differences and contrasts

35.57 The studentized range test of 35.53 leads immediately to simultaneous confidence intervals for all $\frac{1}{2}k(k-1)$ differences between true group means $(\theta_i - \theta_j)$, proposed by Tukey (1951). For whatever the θ_i may be, the random variables $\bar{x}_i - \theta_i$ are identically and independently normally distributed with zero means and variances σ^2/N , and the probability is $1 - \alpha$ that their studentized range will not exceed q_α defined in 35.53. It follows that simultaneously for all $i \neq j$,

$$\text{Prob} \left\{ \left| \frac{(\bar{x}_i - \theta_i) - (\bar{x}_j - \theta_j)}{s/N^{1/2}} \right| \leq q_\alpha \right\} = 1 - \alpha \quad (35.104)$$

so that

$$(\bar{x}_i - \bar{x}_j) - q_\alpha s/N^{1/2} \leq \theta_i - \theta_j \leq (\bar{x}_i - \bar{x}_j) + q_\alpha s/N^{1/2} \quad (35.105)$$

is simultaneously satisfied for all $i \neq j$ with probability $1 - \alpha$.

Exercise 35.11 shows that the method extends to negatively equi-correlated multi-normal \bar{x}_i .

35.58 The method of 35.57 enables us to make simultaneous statements about all $\frac{1}{2}k(k-1)$ differences $\theta_i - \theta_j$ with a known overall confidence coefficient $1 - \alpha$. In many applications of AV, we are interested not only in the differences but also in other linear combinations of the θ_i with constant coefficients summing to zero. Such a linear combination is called a *contrast*, defined by

$$\psi = \sum_{i=1}^k c_i \theta_i, \quad \sum_{i=1}^k c_i = 0. \quad (35.106)$$

The most obviously useful contrast other than the difference between any θ_i and θ_j is the difference between the average of any subset of p of the k parameters θ_i and the average of the $k-p$ others. Interactions (defined in 35.18, 35.41) are also at once seen to be contrasts.

The method of 35.57 is easily adapted so that every contrast, and not merely every

difference, of the θ_i is simultaneously covered by an interval. Since the number of contrasts is infinite, the resulting gain in generality is considerable.

35.59 Write $z_i = \bar{x}_i - \theta_i$, and let $\sum_{i=1}^k c_i = 0$. Consider the maximum possible value of $\sum_i c_i z_i$. Since $\sum c_i = 0$, the sum of the positive c_i is $\frac{1}{2} \sum_i |c_i|$ as is also the sum of the negative c_i . We therefore see that

$$\sum_i c_i z_i \leq \left(\frac{1}{2} \sum_i |c_i| \right) \max_i |z_i - z_j|,$$

i.e.

$$\sum_i c_i (\bar{x}_i - \theta_i) \leq \left(\frac{1}{2} \sum_i |c_i| \right) \max_i |(\bar{x}_i - \theta_i) - (\bar{x}_j - \theta_j)|. \quad (35.107)$$

Referring back to (35.104), we see that (35.107) implies that for any choice of the c_i with $\sum_i c_i = 0$,

$$\text{Prob} \left\{ \left| \frac{\sum_i c_i (\bar{x}_i - \theta_i)}{s/N^{\frac{1}{2}}} \right| \leq \left(\frac{1}{2} \sum_i |c_i| \right) q_\alpha \right\} = 1 - \alpha,$$

and hence that

$$\sum_i c_i \bar{x}_i - \left(\frac{1}{2} \sum_i |c_i| \right) q_\alpha s/N^{\frac{1}{2}} \leq \sum_i c_i \theta_i \leq \sum_i c_i \bar{x}_i + \left(\frac{1}{2} \sum_i |c_i| \right) q_\alpha s/N^{\frac{1}{2}} \quad (35.108)$$

is simultaneously satisfied for all contrasts $\psi = \sum c_i \theta_i$, with probability $1 - \alpha$. The method again generalizes to negatively equi-correlated multinormal \bar{x}_i (cf. Exercise 35.11).

35.60 In 35.59, simultaneous confidence intervals for all contrasts were obtained from intervals for all differences by the use of a rather wasteful inequality. It is not surprising, therefore, that in general these are not the most useful intervals for all contrasts. To obtain a more useful set, we make an entirely different approach.

The estimator of any contrast (35.106) is

$$\hat{\psi} = \sum_i c_i \hat{\theta}_i. \quad (35.109)$$

Clearly,

$$E(\hat{\psi}) = \psi,$$

and, further, if the $\hat{\theta}_i$ are normally distributed, so will $\hat{\psi}$ be. If we now consider any set of r ($\leq k$) estimated contrasts, which we write in the form $\hat{\Psi} = \mathbf{C}\hat{\boldsymbol{\theta}}$, it will be multinormally distributed (cf. 15.4, Vol. 1) with mean vector equal to $\boldsymbol{\psi} = \mathbf{C}\boldsymbol{\theta}$ and dispersion matrix

$$\mathbf{V} = \mathbf{V}(\hat{\Psi}) = \mathbf{C}\mathbf{V}(\hat{\boldsymbol{\theta}})\mathbf{C}'. \quad (35.110)$$

In our present discussion of the one-way classification with equal frequencies N , $\mathbf{V}(\hat{\boldsymbol{\theta}})$ is diagonal, with elements σ^2/N , so that

$$\mathbf{V} = \frac{\sigma^2}{N} \mathbf{C}\mathbf{C}'. \quad (35.111)$$

We assume that \mathbf{V} is non-singular.

35.61 The result of 15.10 now implies that the quadratic form

$$Q = (\hat{\psi} - \psi)' V^{-1} (\hat{\psi} - \psi)$$

has a χ^2 distribution with degrees of freedom equal to r , the rank of V . Independently of Q , the Residual SS (divided by σ^2) is also distributed in this form with, say, ν d.fr. Thus the ratio

$$F = (Q/r)/(s^2/\sigma^2),$$

where $s^2 = (\text{Residual SS})/\nu$ as usual, has the variance-ratio distribution with (r, ν) d.fr. In the simplest case, (35.111), this gives the statistic

$$F = (\hat{\psi} - \psi)' (CC')^{-1} (\hat{\psi} - \psi) / (rs^2/N).$$

If we call the $100(1-\alpha)$ per cent point of this F -distribution $F_{\alpha, r, \nu}$, we now have

$$\text{Prob} \{ (\hat{\psi} - \psi)' (CC')^{-1} (\hat{\psi} - \psi) / (s^2/N) \leq r F_{\alpha, r, \nu} \} = 1 - \alpha. \quad (35.112)$$

The corresponding general result is at once available by using (35.110) instead of (35.111).

35.62 Since V must be non-singular, its rank r cannot exceed q , the number of linearly independent comparisons possible among the θ_i ($k-1$ in the one-way classification—see Example 35.1), which is equal to the d.fr. for their SS in the AV table. (35.112) with $r = q$ holds for *any* set of q linearly independent contrasts we may choose, but this does not imply that it holds for *every* such set simultaneously. However, Scheffé (1953, 1959) showed by geometrical methods that it does imply for every single contrast ψ simultaneously that

$$\text{Prob} \{ (\hat{\psi} - \psi)^2 \leq q F_{\alpha, q, \nu} \hat{V}(\hat{\psi}) \} = 1 - \alpha. \quad (35.113)$$

Here, $\hat{V}(\hat{\psi})$ is the estimated variance of $\hat{\psi}$, in which σ^2 is estimated by s^2 with ν d.fr. Cf. Exercise 35.12 for an analytic proof and Exercise 35.19 for an extremely simple algebraic one.

Scheffé (1953, 1959) went on to show numerically that the intervals for all contrasts yielded by (35.113) are generally shorter than those obtained from (35.108) unless the contrasts happen to be differences—for which (35.108) reduces to (35.105), designed specifically for differences—or otherwise have very few non-zero c_i . Moreover, Scheffé's method is not restricted by the need to have $(\bar{x}_i - \theta_i)$ distributed with equal variances, an assumption fundamental to the argument of 35.57.

35.63 If we now reconsider the variance-ratio test of the overall hypothesis that all the θ_i are equal ((35.22) in the one-way classification case), we see that this hypothesis (cf. (35.20)) states that q linearly independent contrasts are all zero. This implies that *all* contrasts are zero, for every contrast may be regarded as a linear combination of the q linearly independent ones. Thus the overall test is logically equivalent to testing the hypothesis that each of the infinite number of possible contrasts is zero, i.e. seeing whether at least one of the infinite number of intervals given by (35.113) does not cover the value zero. (See also Exercises 35.12 and 35.19.) This property extends at once to Gabriel's (1964) simultaneous test procedure in 35.54: a subset will be adjudged heterogeneous by that method if and only if some contrast within the subset has interval (35.113) not covering zero.

This is the main use of Scheffé's all-contrasts method: once the overall test has rejected the homogeneity hypothesis, the all-contrasts method may be used to examine any contrasts to reveal whether they are in fact the reasons for rejecting the hypothesis, and to calculate confidence intervals for them—they need not be nominated in advance. A natural way of seeking the contrasts which are to "blame" for the rejection of overall homogeneity is to start with all $\frac{1}{2}k(k-1)$ differences. All this may be done without affecting the size of the overall test. If the reader will now refer back to the original discussion of the purposes of multiple comparisons in 35.51, he will probably agree that Scheffé's all-contrasts method is very close to achieving those purposes.

Gabriel (1966) gives a general theory of simultaneous test procedures.

35.64 Dunn (1961) considers a procedure intermediate between setting confidence intervals for a single contrast and setting them for all contrasts. Her method requires the prior nomination of m contrasts as of special interest. The intervals obtained (based on "Student's" t -statistic) are shorter than those obtained from either Tukey's or Scheffé's method if k (the number of parameters) exceeds 2 and m is not too large—this advantage increases as k , or the number of d.fr. for Residual, or the confidence coefficient $1-\alpha$, increases. The very simple result which underlies this method is given in Exercise 35.14. The procedure is improved by Siotani (1964).

Ordered and metrical classifications

35.65 Throughout this chapter, the classification variables have been quite general, no assumption having been made about whether, e.g., the groups in a classification are ordered in any way. However, if precise information is available concerning the basis of the classification, the SS in the AV table can be further partitioned into corresponding components. For example, if it is known that the groups correspond to equally-spaced values of an underlying variable, the orthogonal polynomials discussed in 28.18–20, Vol. 2, may be used to assign a single d.fr. to the linear, quadratic, cubic, and higher-degree effects of the classifying variable, if necessary proceeding until all $(k-1)$ d.fr. are exhausted. The method used is precisely that of Example 28.3.

In more complex classifications, interactions as well as row- (or other) effects may be partitioned in this way if all the underlying variables are equally spaced. Computational methods are given by R. L. Anderson and Bancroft (1952) and in Fisher and Yates' *Tables*.

35.66 Bartholomew (1961) considers the case of an ordered classification as an alternative to the hypothesis of homogeneity of the groups. This is precisely the situation discussed in 31.74. When the n_i are all equal, the distribution-free test based on (31.151) is found to have higher power asymptotically than the LR test if the θ_i are equally spaced, but to be less powerful if, at the other extreme, all the θ_i are equal except one. See Exercise 35.15, and also the paper by Chacko (1963).

Analysis of covariance

35.67 A natural extension of AV arises when, in an analysis of classified data such as we have been discussing in this chapter, we have available to us not only the

observations on y but also the values of one or more further variables x , known or suspected to influence the value of y . If the data were not classified, we should carry out here an ordinary regression analysis of y on the x 's, but what we wish to investigate now is the joint effect upon y of the classification (possibly a complex one) and of the measured variables x . There is more than one possible purpose for such an analysis.

Commonly, the values of the x 's are to be eliminated by regression methods, so that we may be free to analyse the effect of the classification upon y after discounting the influence of x —for example, when x is the value on an earlier occasion (before the treatments giving rise to the classification) of y itself. Thus, if the effect of different teaching methods upon children's performance in a school subject is to be measured, their initial levels of performance would be the values of x , and their final levels the values of y ; an alternative would be to take a measure of general intelligence as the value of x in the analysis; or it might be thought worth while to include both the initial level of performance and the intelligence measure as x -variables. We may describe this as an elimination-motivated analysis. Its methods are to ensure "fair" comparisons among treatments and also, by removing unwanted variation due to x , the reduction of residual variation.

It will obviously make interpretation of results simpler if, as in our example, x is unaffected by the treatment yielding the classification, but the analysis may be carried out in any case.

However, we may, alternatively or additionally, be interested in whether the regression of y on x is affected by the classification at all; in our example, we might ask whether the regression of final performance-level upon initial performance-level is the same whichever method is used. Our motive for the analysis is not now elimination, but intrinsic interest in the relationship between the variables.

35.68 This branch of the subject has come to be called *Analysis of Covariance*, because the regression calculations involve partitioning sums of products of y and x in the same way as ordinary AV involves the partitioning of SS. The variables x are usually called *concomitant variables*, implying that y is of prior interest.

An extended expository review of uses of the Analysis of Covariance is contained in a set of seven papers in the September 1957 issue of *Biometrics* (Vol. 13, No. 3). A clear introductory account is given by D. R. Cox (1958). We shall be concerned purely with its theoretical aspects.

35.69 Since we have already seen that regression analysis and AV can each be treated within the framework of a linear model, it is evident that Analysis of Covariance, a mixture of the two, can be so treated. The interpretative convenience of having the concomitant variables unaffected by the treatments is now seen to be a special case of the convenience of having different sets of regressors uncorrelated in linear models. One can therefore set up a linear model *ab initio* for any situation requiring an Analysis of Covariance.

However, this tedious process can be avoided if the AV which is (so to speak) embedded within the Analysis of Covariance is of a known form; we can then extend an AV to produce an associated Analysis of Covariance, by extending the linear model

appropriately. Moreover, it transpires that this process of extension is a quite general one, enabling us to introduce extra parameters into an existing linear model. As its heading indicates, the algebraic discussion which follows has, therefore, no limitation to AV situations.

The extension of a linear model to include further parameters

35.70 Suppose that a linear model (singular or non-singular)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad (35.114)$$

has been fitted to the observations, and that the LS estimator of $\boldsymbol{\theta}$ obtained by the methods of Chapter 19 is

$$\hat{\boldsymbol{\theta}} = \mathbf{T}\mathbf{y}, \quad (35.115)$$

where of course $\mathbf{T} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ in the non-singular case. We find that the Residual SS is

$$\begin{aligned} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}) &= \{(\mathbf{I} - \mathbf{X}\mathbf{T})\mathbf{y}\}'\{(\mathbf{I} - \mathbf{X}\mathbf{T})\mathbf{y}\} \\ &= \mathbf{y}'(\mathbf{I} - \mathbf{X}\mathbf{T})\mathbf{y}, \end{aligned} \quad (35.116)$$

since $\mathbf{TX} = \mathbf{I}$ is the condition for unbiasedness of $\hat{\boldsymbol{\theta}}$. The matrix $(\mathbf{I} - \mathbf{X}\mathbf{T})$ is idempotent.

35.71 Now consider the extended model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (35.117)$$

or

$$\mathbf{y} - \mathbf{Z}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}.$$

If $\boldsymbol{\beta}$ were known, this would have the solution, from (35.114-15),

$$\hat{\boldsymbol{\theta}} = \mathbf{T}(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}), \quad (35.118)$$

and thus the LS solution of (35.117) for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ may be obtained by solving for $\hat{\boldsymbol{\beta}}$ alone

$$\mathbf{y} = \mathbf{X}\mathbf{T}(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

or

$$(\mathbf{I} - \mathbf{X}\mathbf{T})\mathbf{y} = (\mathbf{I} - \mathbf{X}\mathbf{T})\mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \quad (35.119)$$

(35.119) is a linear model, and we assume that it is non-singular. We therefore have, from Chapter 19,

$$\hat{\boldsymbol{\beta}} = \{\mathbf{Z}'(\mathbf{I} - \mathbf{X}\mathbf{T})\mathbf{Z}\}^{-1}\mathbf{Z}'(\mathbf{I} - \mathbf{X}\mathbf{T})\mathbf{y}, \quad (35.120)$$

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) = \sigma^2 \{\mathbf{Z}'(\mathbf{I} - \mathbf{X}\mathbf{T})\mathbf{Z}\}^{-1}. \quad (35.121)$$

35.72 The reduction in the Residual SS due to the extension of the model is

$$\{(\mathbf{I} - \mathbf{X}\mathbf{T})\mathbf{Z}\hat{\boldsymbol{\beta}}\}'\{(\mathbf{I} - \mathbf{X}\mathbf{T})\mathbf{Z}\hat{\boldsymbol{\beta}}\} = \hat{\boldsymbol{\beta}}'\mathbf{Z}'(\mathbf{I} - \mathbf{X}\mathbf{T})\mathbf{y}. \quad (35.122)$$

On comparing this with (35.116), we see that they embody the same matrix $(\mathbf{I} - \mathbf{X}\mathbf{T})$; (35.122) differs only in that $\hat{\boldsymbol{\beta}}\mathbf{Z}'$ replaces \mathbf{y}' in premultiplying this matrix. This simplifies the computation of (35.122), since we have to replace a quadratic form in \mathbf{y} by a set of corresponding bilinear forms, obtained by each column of \mathbf{Z} in turn replacing \mathbf{y}' in (35.116). These bilinear forms, assembled into a column vector, are premultiplied by $\hat{\boldsymbol{\beta}}$ to obtain (35.122).

The difference between (35.116) and (35.122),
 $(\mathbf{y} - \mathbf{Z}\hat{\beta})'(\mathbf{I} - \mathbf{X}\mathbf{T})\mathbf{y}$,

is the Residual SS when the extended model (35.117) has been fitted. It is easy to see that a reduction of exactly the same form as (35.122) applies to the minimized SS under any constraints upon the elements of θ : the only change is that $(\mathbf{I} - \mathbf{X}\mathbf{T})$ is replaced by the matrix \mathbf{Q} of the quadratic form of the minimized SS. The analogues of (35.122-3) are then $\beta' \mathbf{Z}' \mathbf{Q} \mathbf{y}$ and $(\mathbf{y} - \mathbf{Z}\beta)' \mathbf{Q} \mathbf{y}$ respectively.

35.73 The application of the results of 35.70-2 to Analysis of Covariance is immediate. Corresponding to the SS which emerge in AV situations, (35.122) produces sums of products. Computational instructions and examples are given by Scheffé (1959), and worked examples by R. L. Anderson and Bancroft (1952) and in the September 1957 *Biometrics* issue.

EXERCISES

35.1 Verify that if all $n_{ij} = m$, the SS in (35.63-4) become

$$S_4 = (\sum_i \sum_j \sum_p y_{ijp})^2 / (rcm),$$

$$S_1 = \sum_i (\sum_j \sum_p y_{ijp})^2 / (cm) - S_4,$$

$$S_2 = \sum_j (\sum_i \sum_p y_{ijp})^2 / (rm) - S_4,$$

$$S_3 = \sum_i \sum_j (\sum_p y_{ijp})^2 / m - (S_1 + S_2 + S_4),$$

$$S_R = \sum_i \sum_j \sum_p y_{ijp}^2 - (S_1 + S_2 + S_3 + S_4),$$

and show that if $m = 1$, these SS reduce (the summation over p now being redundant) to

$$S_4 = (\sum_i \sum_j y_{ij})^2 / (rc),$$

$$S_1 = \sum_i (\sum_j y_{ij})^2 / c - S_4,$$

$$S_2 = \sum_j (\sum_i y_{ij})^2 / r - S_4,$$

$$S_3 = \sum_i \sum_j y_{ij}^2 - (S_1 + S_2 + S_4),$$

$$S_R \equiv 0.$$

35.2 Verify that if all $n_{ij} = m$, the matrix \mathbf{C} defined at (35.39) becomes

$$\mathbf{C}_{(r-1)(c-1) \times (r-1)(c-1)} = m \begin{pmatrix} 2\mathbf{E} & \mathbf{E} & \mathbf{E} & \cdot & \cdot & \cdot & \mathbf{E} \\ \mathbf{E} & 2\mathbf{E} & \mathbf{E} & \cdot & \cdot & \cdot & \mathbf{E} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathbf{E} & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{E} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{E} & 2\mathbf{E} \end{pmatrix}$$

where

$$\mathbf{E}_{(c-1) \times (c-1)} = \begin{pmatrix} 2 & 1 & \cdot & \cdot & \cdot & 1 \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ 1 & \cdot & \cdot & \cdot & 1 & 2 \end{pmatrix}$$

Show that

$$\mathbf{C}^{-1} = \frac{1}{rm} \begin{pmatrix} (r-1)\mathbf{E}^{-1} & -\mathbf{E}^{-1} & \cdot & \cdot & \cdot & -\mathbf{E}^{-1} \\ -\mathbf{E}^{-1} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & -\mathbf{E}^{-1} \\ -\mathbf{E}^{-1} & \cdot & \cdot & \cdot & -\mathbf{E}^{-1} & (r-1)\mathbf{E}^{-1} \end{pmatrix},$$

where

$$\mathbf{E}^{-1} = \frac{1}{c} \begin{pmatrix} c-1 & -1 & \cdot & \cdot & \cdot & -1 \\ -1 & c-1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & -1 \\ -1 & \cdot & \cdot & -1 & c-1 \end{pmatrix},$$

and hence verify that the LS estimator of θ_{ij} is given by (35.50).

35.3 Generalize the AV table of Example 35.5 to a three-way hierarchical classification and give the three LR tests of the hypothesis that there is no variation in (a) the group means; (b) sub-group means within groups; and (c) sub-sub-group means within sub-groups.

35.4 In Example 35.4, show that if it is postulated that H_3 of (35.54) holds, so that there are known to be no interactions, the SS attributable to row-effects is $M - S_2$, and the SS attributable to column-effects is $M - S_1$, where S_1 and S_2 are defined by (35.59-60). Show that the Residual MS has $(n - r - c + 1)$ d.fr. in this case.

35.5 Show that if, in a $(r \times 2)$ cross-classification, the differences of cell-means $d_i = y_{i1} - y_{i2}$, are analysed by the method of weighted squares of means applied to the \bar{y}_i in 35.31, the SS $\sum_i W_i (d_i - \sum_i W_i d_i / \sum_i W_i)^2$, where $W_i = n_{i1}^{-1} + n_{i2}^{-1}$, provides the test of the hypothesis that the interactions are all zero.

(Yates, 1934)

35.6 In a $2 \times 2 \times 2 \times \dots = 2^m$ cross-classification, show that if a $2 \times 2^{m-1}$ table is formed for any one of the classifications (A , say) against all possible combinations of the others, the unweighted mean of the differences of the row-cell means provides an efficient estimator and test of the effect of A . Show that this is a generalization of 35.31, and that any interaction can similarly be tested.

(Yates, 1934)

35.7 The table below gives Brandt's data, used by Yates (1934), for an 8×2 cross-classification by breed and sex of 533 slaughtered pigs; cell frequencies n_{ij} and the cell totals of the variable studied (the logarithm of the percentage bacon yielded by the carcass) are shown overleaf.

THE ADVANCED THEORY OF STATISTICS

64

Breed	Female		Male	
	n_{i1}	$\sum y_{i1p}$	n_{i2}	$\sum y_{i2p}$
I	33	66.55	89	181.04
II	51	98.69	141	281.43
III	13	25.90	17	34.20
IV	4	7.62	9	17.58
V	8	14.64	4	8.20
VI	15	28.11	32	64.42
VII	35	66.90	47	90.52
VIII	12	23.32	23	46.70
TOTALS	171	331.73	362	724.09

The Total SS, which is not obtainable from the above, was 13.0142. Using 35.31 and Exercise 35.5, show that (neglecting the 1 d.fr. for the general mean) the AV is

Variation	D.fr.	SS	MS
Due to breeds	7	0.6056	0.0865
Due to sexes	1	0.3032	0.3032
Breeds and sexes	8	1.0415	0.0329
Interactions	7	0.2300	0.0329
Total between classes	15	1.2715	0.0848
Residual	517	11.7427	0.0227
Total minus general mean	532	13.0142	

Show that breed and sex each have effects upon bacon yield, but that they do not interact.

35.8 If there is only one observation per cell in Example 35.6, show that the AV table (35.95) holds good with the Residual $SS \equiv 0$, and that if the second-order interactions are dropped from the linear model, the SS with $(r-1)(c-1)(l-1)$ d.fr. becomes the Residual SS for testing all other parameters (cf. Example 35.3).

35.9 Show that the AV for a $(r \times c)$ cross-classification with m observations in every cell may be formally constructed from a $(r \times c \times l)$ cross-classification with exactly one observation per cell, in which the layer classification factor is "replication," its main effect and all interactions concerned with it being defined to be identically zero.

35.10 For the Newman-Keuls procedure defined in 35.54, show that if the true means of any subset of p of the k groups are all equal, while all the other groups have different means, the probability of wrongly adjudging a pair in the "cluster" to be heterogeneous cannot exceed α .

If there are m such clusters of truly equal means, show that the probability of wrongly adjudging a pair from any such cluster to be heterogeneous cannot exceed $m\alpha$.

(Hartley, 1955)

35.11 If in 35.57 the k variables $(\bar{x}_i - \theta_i)$ are multinormally distributed with all variances σ^2/N and all covariances equal to $-\lambda^2 \sigma^2/N$, show that the k variables $z_i = \bar{x}_i - \theta_i + x_0$ are independently normal if x_0 is a normal variable independent of the \bar{x}_i with zero mean and variance $\lambda^2 \sigma^2/N$. By applying the method of 35.57 to the z_i , show that, with probability $1 - \alpha$,

$$\bar{x}_i - \bar{x}_j - q_\alpha \{(1 + \lambda^2) s^2 / N\}^{\frac{1}{2}} \leq \theta_i - \theta_j \leq \bar{x}_i - \bar{x}_j + q_\alpha \{(1 + \lambda^2) s^2 / N\}^{\frac{1}{2}}$$

holds simultaneously for all $\frac{1}{2}k(k-1)$ differences $\theta_i - \theta_j$. Show that the result of 35.59 generalizes in exactly the same way. (Scheffé, 1959)

35.12 In the one-way classification of Example 35.1, without loss of generality, take $\sum_{i=1}^k n_i \bar{\theta}_i$ as origin and σ as unit. Show that the value of $t^2 = (\sum_i c_i \bar{\theta}_i)^2 / (\sum_i c_i^2 / n_i)$ is maximized for choice of the c_i when $c_i \propto n_i \bar{\theta}_i$, so that $\sum_i c_i = 0$ and $|t|$ is the largest observed absolute ratio of a contrast to its standard error. Show further that $t^2 = S_2$, the numerator SS (defined by (35.21)) of the overall variance-ratio test defined by (35.22), so that the overall test essentially tests the largest observed contrast.

(This result holds quite generally—
cf. Scheffé (1959) and Gabriel (1964))

35.13 By defining a dummy parameter $\theta_0 \equiv 0$ with estimator $\hat{\theta} \equiv 0$ also, show that all linear combinations $\psi = \sum_i c_i \theta_i$ (where $\sum_i c_i$ need not be zero) may have confidence intervals set for them by (35.113) with q increased by 1; and that the method of (35.108) may similarly be used if k is increased by 1 and $\frac{1}{2} \sum_i |c_i|$ is replaced by

$$\max \left\{ \sum_i c_i, c_i > 0; \left| \sum_i c_i \right|, c_i < 0 \right\}$$

(Tukey, 1953; Scheffé, 1959)

35.14 In 35.64, consider k non-independent events with equal probabilities P_1 of occurring. Show that P_k , the probability that they all occur, satisfies $P_k \geq 1 - k(1 - P_1)$, and let P_1 be the probability $1 - \alpha$ that a "Student's" t -statistic (ν d.f.) lies in the interval $(-t_\alpha, t_\alpha)$. Show that if m linear combinations $\lambda_s = \sum_{i=1}^k c_{si} \theta_i$, $s = 1, 2, \dots, m$, are estimated by $l_s = \sum_i c_{si} \hat{\theta}_i$, then $(l_s - \lambda_s) / \{\hat{V}(l_s)\}^{1/2}$ is distributed in "Student's" form with ν d.f. for each s . Hence show that

$$\text{Prob } \{l_s - t_\alpha [\hat{V}(l_s)]^{1/2} \leq \lambda_s \leq l_s + t_\alpha [\hat{V}(l_s)]^{1/2}\} \geq 1 - \alpha.$$

(Dunn, 1961)

35.15 Consider the distribution-free statistic U for testing k samples against ordered alternatives defined at (31.151), and the competitive statistic U' defined similarly, except that U_{pq} is replaced by

$$U'_{pq} = \sum_{i=1}^{n_p} \sum_{j=1}^{n_q} (x_{pi} - x_{qj}) \equiv n_p n_q (\bar{x}_p - \bar{x}_q),$$

so that

$$U' = \sum_{p=1}^k \sum_{q=p+1}^k n_p n_q (\bar{x}_p - \bar{x}_q).$$

For normally distributed observations x_{is} with means $E(x_{is}) = \theta_i$ and equal variances σ^2 , show that if $n_1 = n_2 = \dots = n_k = N$, the asymptotic power functions of U and U' in testing equality of the θ_i against the alternative hypothesis that the θ_i are equally spaced, are respectively

$$G\{\Delta(3/\pi)^{1/2} - \lambda_\alpha\} \quad \text{and} \quad G\{\Delta - \lambda_\alpha\},$$

where G is the standardized normal d.f., $\Delta^2 = \sum_{i=1}^k (\theta_i - \bar{\theta})^2 / \sigma^2$, and $G\{-\lambda_\alpha\} = \alpha$ as in Chapter 25.

Deduce that the ARE of U compared to U' is $3/\pi$. Show that this last result holds whatever the relations among the ordered θ_i .

(Bartholomew, 1961)

THE ADVANCED THEORY OF STATISTICS

56

35.16 In 35.54, let R and P be any subsets of the k groups such that R contains P . Show that (35.21) calculated for R can never be less than the same SS calculated for P . Show

35.17 Show that the simultaneous test procedure of 35.54 has the property that the probability of wrongly adjudging any homogeneous subset to be heterogeneous is at most α . (Gabriel, 1964)

35.18 In 35.54 define a new step-by-step procedure based on the SS (35.21), but applied in the manner of the last paragraph of 35.53. This has the same overall size α as the simultaneous test procedure of 35.54. Show that for such a step-by-step procedure the critical value used for the SS must increase with the size of the subset being tested, and hence that a subset can only be adjudged heterogeneous by the step-by-step method if it is by the simultaneous method. (Gabriel, 1964)

35.19 In 35.62-3, show by the Cauchy inequality that

$$\max_{c_i} \left\{ \sum_{i=1}^k c_i (\hat{\theta}_i - \theta_i) \right\}^2 = \sum_{i=1}^k c_i^2 \sum_{i=1}^k (\hat{\theta}_i - \theta_i)^2,$$

and hence that the squared difference between the largest observed contrast and its expectation is distributed as $(\sum_i c_i^2) q s^2 F$, where F is the test statistic for the overall hypothesis that the θ_i are equal. Hence establish (35.113).

(This proof is due to M. H. Belz and A. M. W. Verhagen)

OTHER MODELS FOR THE ANALYSIS OF VARIANCE

36.1 Throughout the previous chapter, we have been concerned with the application of the general linear model to the analysis of observations classified into groups. Underlying the whole of the discussion was the assumption, explicitly written into the linear model, that the classification affected the observations through their mean values, but not otherwise. In the case of the general linear model, therefore, Analysis of Variance (AV) is accurately described as an analysis of *means*, which is carried out through certain sums of squares (SS) computed from the observations.

It is a remarkable fact that we are led to very similar (and in the simpler cases, even identical) computations of SS when investigating a quite different type of situation. In the early development of the subject, this similarity in analysis tended to obscure the fundamental distinctions between the underlying mathematical models, which were first set out in detail by Eisenhart (1947). AV based on the LS analysis of the general linear model, as in Chapter 35, was there called Model I AV, a name in common use subsequently. The other well-established mathematical model, Model II AV, will now be investigated.

The reader will realize that the Model I definition (cf. 35.4, 35.9–10) of the term “Analysis of Variance” must now be broadened. We define AV generally as the study, whether by means of classified data (cf. 35.15) or not, of the resolution of SS into component SS attributable to various factors, acting singly and in combination.

Model II: components of variance

36.2 Instead of the general linear model (19.8), consider the superficially similar model

$$\underset{(n \times 1)}{\mathbf{y}} = \underset{(n \times 1)}{1\theta} + \underset{(n \times p)(p \times 1)}{\mathbf{X}\mathbf{u}} + \underset{(n \times 1)}{\boldsymbol{\epsilon}}. \quad (36.1)$$

In (36.1), as in Chapter 35 (and earlier in Exercise 19.1, Vol. 2), we isolate the “general mean” θ , which here will need no subscript. As before, $\mathbf{1}$ is a vector of units and \mathbf{X} a known matrix of constants, while $\boldsymbol{\epsilon}$ is the vector of errors in the observations. The crucial change is the replacement of the parameter-vector $\boldsymbol{\theta}$ in (19.8) by a vector \mathbf{u} of p random variables. Thus (36.1) states that y_i ($i = 1, 2, \dots, n$) is composed of the general mean θ , plus a linear combination of p random variables u_j , plus an error term, ϵ_i . There are $(p+1)$ random components of y_i , instead of only the one in (19.8).

We assume, as at (19.9–10) for the general linear model, that

$$E(\boldsymbol{\epsilon}) = \mathbf{0}; \quad V(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma_{\epsilon}^2 \mathbf{I}, \quad (36.2)$$

where we have added an identifying subscript to the error variance σ_{ϵ}^2 to distinguish it from the variances of our new random variables. We further assume that

$$E(\mathbf{u}) = \mathbf{0}, \quad (36.3)$$

which is why we isolated the general mean θ , and that

$$\left. \begin{aligned} \text{var } u_j &= \sigma_j^2; \\ \text{cov}(u_j, u_l) &= 0, \quad l \neq j; \end{aligned} \right\} \quad (36.4)$$

Thus all our random variables have zero means and are uncorrelated in pairs.

36.3 The parameters of interest in the model (36.1) are the σ_j^2 , the variances of the u_j , and the error variance σ_ϵ^2 . We may reduce the dimensionality of the problem by taking account of the fact that some of the σ_j^2 may be known to be equal. Suppose, then, that there are k different variances σ_j^2 , where $k \leq p$. We rewrite (36.1) in the form

$$y = \theta + \sum_{j=1}^k \mathbf{X}_j u_j + \epsilon, \quad (36.5)$$

where the \mathbf{u}_j are now vectors (subvectors of \mathbf{u}), which contain p_j uncorrelated random variables $\left(\sum_{j=1}^k p_j = p \right)$, with zero means and variance σ_j^2 . \mathbf{X}_j is a $(n \times p_j)$ submatrix of \mathbf{X} .

In (36.5) we write \mathbf{W} for the matrix $E(\mathbf{y}\mathbf{y}')$ and \mathbf{V}_y for the dispersion matrix of y , assumed to be non-singular, so that, using (36.2-4),

$$\mathbf{V}_y = \mathbf{W} - \theta^2 \mathbf{1}\mathbf{1}' = \sum_{j=1}^k \sigma_j^2 \mathbf{X}_j \mathbf{X}_j' + \sigma_\epsilon^2 \mathbf{I}. \quad (36.6)$$

Essentially, therefore, we wish to estimate the parameters which appear as coefficients of the $(k+1)$ components of the dispersion matrix of the observations, and it is appropriate that (36.5) is often called a *components of variance* model for AV as an alternative to the "Model II" label.

(36.6) emphasizes an essential distinction between (36.5) and the general linear model: the observations are not now all uncorrelated (\mathbf{V}_y is not diagonal), since they have as components linear functions of the same variables \mathbf{u}_j .

Example 36.1

Consider the model (36.5) with $k = 1$, \mathbf{X}_1 a $(n \times p)$ matrix patterned like \mathbf{X} in Example 35.1, and \mathbf{u}_1 a $(p \times 1)$ vector. This is a one-way classification problem for the present model, in which the observations are in p groups and the q th observation in the i th group is

$$y_{iq} = \theta + u_i + \epsilon_{iq},$$

where the u_i (the p components of \mathbf{u}_1) and the ϵ_{iq} all have zero means, are all uncorrelated, and have variances

$$\begin{aligned} \text{var } u_i &= \sigma_1^2, \quad \text{all } i, \\ \text{var } \epsilon_{iq} &= \sigma_\epsilon^2, \quad \text{all } i, q. \end{aligned}$$

It follows immediately that

$$\text{var } y_{iq} = \sigma_1^2 + \sigma_\epsilon^2,$$

so that here the two parameters of the model are literally components of the variance of the observations. It was this which gave rise to the term "components of variance" which we have attached to the model (36.5) in general.

36.4 Our investigation of the properties of the model (36.5) must start from the beginning; none of the LS theory used in the general linear model is now applicable. Our treatment follows that of Graybill and Hultquist (1961).

(36.5) can be written more symmetrically as

$$\mathbf{y} = \sum_{j=0}^{k+1} \mathbf{X}_j \mathbf{u}_j, \quad (36.7)$$

where we define

$$\mathbf{X}_0 \equiv \mathbf{1}, \quad \mathbf{u}_0 \equiv \theta \text{ (scalar)}, \quad \mathbf{X}_{k+1} \equiv \mathbf{I}, \quad \mathbf{u}_{k+1} = \boldsymbol{\epsilon}.$$

There are $(k+2)$ parameters in (36.7), namely θ and the $(k+1)$ variances appearing in the model. The assumptions (36.2-4) are now summarized as

$$\left. \begin{aligned} E(\mathbf{u}_j) &= \mathbf{0}, \quad V(\mathbf{u}_j) = E(\mathbf{u}_j \mathbf{u}_j') = \sigma_j^2 \mathbf{I}, \\ E(\mathbf{u}_j \mathbf{u}_i') &= \mathbf{0}, \quad i \neq j; \quad i, j = 1, 2, \dots, k+1, \end{aligned} \right\} \quad (36.8)$$

where $\sigma_{k+1}^2 \equiv \sigma_{\epsilon}^2$. Formally, we may write

$$\sigma_0^2 = E(\mathbf{u}_0 \mathbf{u}_0') = \theta^2, \quad (36.9)$$

so that our parameters are σ_j^2 ($j = 0, 1, 2, \dots, k+1$). We can now rewrite (36.6) as

$$\mathbf{W} = E(\mathbf{y} \mathbf{y}') = \sum_{j=0}^{k+1} \sigma_j^2 \mathbf{A}_j, \quad (36.10)$$

where $\mathbf{A}_j = \mathbf{X}_j \mathbf{X}_j'$, and

$$\mathbf{V}_y = \sum_{j=1}^{k+1} \sigma_j^2 \mathbf{A}_j. \quad (36.11)$$

Unbiased quadratic estimation of the parameters

36.5 In our investigation of the general linear model, we found at (19.19) the condition which must be satisfied if linear functions of the parameters are to be unbiasedly estimated by linear functions of the observations. In the components of variance model (36.7), the parameters occur in (36.10) as coefficients in \mathbf{W} , the matrix of second-order moments of the observations, and it is natural to seek quadratic estimators of them. We now prove that a necessary and sufficient condition that the σ_s^2 be unbiasedly estimable by quadratic forms $\mathbf{y}' \mathbf{C}_s \mathbf{y}$ is that the matrices \mathbf{A}_s are linearly independent.

36.6 First, assume that there exist matrices \mathbf{C}_s such that

$$E(\mathbf{y}' \mathbf{C}_s \mathbf{y}) = \sigma_s^2. \quad (36.12)$$

Using (36.7-8), this implies that

$$\sigma_s^2 = E\left\{ \left(\sum_j \mathbf{X}_j \mathbf{u}_j \right)' \mathbf{C}_s \left(\sum_j \mathbf{X}_j \mathbf{u}_j \right) \right\} = E\left\{ \sum_j \mathbf{u}_j' \mathbf{X}_j' \mathbf{C}_s \mathbf{X}_j \mathbf{u}_j \right\}. \quad (36.13)$$

As at (19.38),

$$E(\mathbf{u}_j' \mathbf{B} \mathbf{u}_j) = \sigma_j^2 \text{tr } \mathbf{B}, \quad j = 1, 2, \dots, k+1, \quad (36.14)$$

and this also holds trivially for $j = 0$, so (36.13) becomes

$$\begin{aligned} \sigma_s^2 &= \sum_{j=0}^{k+1} \sigma_j^2 \text{tr} (\mathbf{X}_j' \mathbf{C}_s \mathbf{X}_j) = \sum_j \sigma_j^2 \text{tr} (\mathbf{X}_j \mathbf{X}_j' \mathbf{C}_s) \\ &= \sum \sigma_j^2 \text{tr} (\mathbf{A}_j \mathbf{C}_s). \end{aligned} \quad (36.15)$$

Equating coefficients of σ_j^2 in (36.15), we find

$$\left. \begin{aligned} \text{tr}(\mathbf{A}_s \mathbf{C}_s) &= 1, \\ \text{tr}(\mathbf{A}_j \mathbf{C}_s) &= 0, \quad j \neq s. \end{aligned} \right\} \quad (36.17)$$

Now let l_0, l_1, \dots, l_{k+1} be any constants satisfying

$$\sum_{j=0}^{k+1} l_j \mathbf{A}_j = \mathbf{0},$$

which, if the l_j were non-zero, would make the \mathbf{A}_j linearly dependent. Since, by (36.16-17), it would then follow that

$$l_s = \sum_j l_j \text{tr}(\mathbf{A}_j \mathbf{C}_s) = \text{tr}(\mathbf{C}_s \sum_j l_j \mathbf{A}_j) = 0,$$

we see that (36.17) can only hold if the l_j are all equal to zero. Thus (36.12) implies linear independence of the \mathbf{A}_s .

36.7 To prove the converse result, let the $\frac{1}{2}n(n+1)$ elements on and above the leading diagonal of the $(n \times n)$ matrices \mathbf{W} and \mathbf{A}_j in (36.10) be placed in the same (arbitrary) order into vectors $\mathbf{w}^*, \mathbf{a}_j^*, j = 0, 1, \dots, k+1$. (36.10) then implies that

$$E(\mathbf{w}^*) = \sum_j \sigma_j^2 \mathbf{a}_j^*. \quad (36.18)$$

If we write \mathbf{A}^* for the $[\frac{1}{2}n(n+1) \times (k+2)]$ matrix formed by juxtaposing the \mathbf{a}_j^* , and σ^2 for the vector of the σ_j^2 , (36.18) becomes

$$E(\mathbf{w}^*) = \mathbf{A}^* \sigma^2. \quad (36.19)$$

Since the \mathbf{A}_j are now assumed linearly independent, so are the \mathbf{a}_j^* which were formed from their elements. Thus \mathbf{A}^* , formed from the \mathbf{a}_j^* , has rank $(k+2)$, and we can find $(k+2)$ linearly independent rows in it which form a non-singular submatrix \mathbf{A}^{**} . We write \mathbf{w}^{**} for the corresponding subvector of \mathbf{w}^* . (36.19) implies that

$$E(\mathbf{w}^{**}) = \mathbf{A}^{**} \sigma^2$$

and hence

$$(\mathbf{A}^{**})^{-1} E(\mathbf{w}^{**}) = \sigma^2,$$

or

$$E\{(\mathbf{A}^{**})^{-1} \mathbf{w}^{**}\} = \sigma^2, \quad (36.20)$$

which establishes that there is a quadratic unbiased estimator of the parameter-vector. We henceforth assume the linear independence of the \mathbf{A}_s .

Sufficient statistics in the normal commutative case

36.8 So far, no assumptions have been made concerning the distributional forms of the random variables u_j in our model. We now investigate the case where each $u_j (j = 1, 2, \dots, k+1)$ is multinormally distributed. Together with the zero means and covariances assumed in (36.8), this implies that the p random variables $u_j, j = 1, 2, \dots, p$, are independent normal variables with zero means.

The normality assumption alone will not take us very much further: to make substantial progress, we must impose the additional condition of commutativity

$$\mathbf{A}_j \mathbf{A}_i = \mathbf{A}_i \mathbf{A}_j, \quad i, j = 0, 1, 2, \dots, k+1. \quad (36.21)$$

The \mathbf{A}_j are symmetric, so we always have

$$\mathbf{A}_j \mathbf{A}_i = \mathbf{A}_j' \mathbf{A}_i' = (\mathbf{A}_i \mathbf{A}_j)'$$

Thus what (36.21) requires is that each $A_i A_j$, as well as each A_j , be symmetric. That this condition is restrictive is shown in Example 36.2.

Example 36.2

Consider again the one-way classification situation in Example 36.1, with $k = 1$. Here $A_0 = \mathbf{1}\mathbf{1}'$ as always, and from Example 35.1,

$$A_1 = \begin{pmatrix} (\mathbf{1}\mathbf{1}')_{n_1} & & 0 \\ & (\mathbf{1}\mathbf{1}')_{n_2} & \\ 0 & & (\mathbf{1}\mathbf{1}')_{n_p} \end{pmatrix}$$

where the suffixes give the number of rows and columns in the submatrices. Multiplication shows that $A_0 A_1$ in its first n_1 columns has every element equal to n_1 , in its next n_2 columns has every element equal to n_2 , and so on until in its last n_p columns every element is equal to n_p . Thus $A_0 A_1$ is symmetric (A_0 and A_1 commute) only if all the n_i are equal. Since $X_2 = \mathbf{I}$, $A_2 = \mathbf{I}$ also and always commutes. The present model will therefore cover the one-way classification only in the balanced case, with equal frequencies in the p groups. Contrast Example 35.1 for Model I, where group frequencies were quite unimportant.

36.9 As Example 36.2 indicates, we can only expect (36.21) to hold in general for the balanced case. We now proceed with our investigation, bearing this restriction in mind.

The normality and independence of the p random variables u_j , $j = 1, 2, \dots, p$, implies that the (correlated) variables y_1, y_2, \dots, y_n are multinormally distributed with

$$E(\mathbf{y}) = \mathbf{X}_0 \mathbf{u}_0 = \mathbf{1}\theta$$

and dispersion matrix \mathbf{V}_y given by (36.11). The quadratic form in the exponent of their multinormal distribution is therefore

$$Q = (\mathbf{y} - \mathbf{1}\theta)' \mathbf{V}_y^{-1} (\mathbf{y} - \mathbf{1}\theta), \quad (36.22)$$

distributed in the chi-squared form with n d.fr. by 15.10.

36.10 Now, because of the commutativity condition (36.21), there exists an orthogonal matrix \mathbf{P} which simultaneously diagonalizes all the A_j so that

$$\mathbf{P} A_j \mathbf{P}' = \mathbf{D}_j \quad (36.23)$$

where \mathbf{D}_j is a diagonal matrix. Moreover, we may choose \mathbf{P} so that one row (say, its first, denoted by \mathbf{P}_1) has elements all equal to $n^{-1/2}$ and

$$\mathbf{P}_1 \mathbf{1} = n^{1/2}, \quad \mathbf{P}_j \mathbf{1} = 0, \quad j \neq 1, \quad (36.24)$$

where \mathbf{P}_j is any set of rows of \mathbf{P} not including \mathbf{P}_1 .

(36.10–11) show that \mathbf{W} and \mathbf{V}_y are also diagonalized by \mathbf{P} , the leading diagonal of $\mathbf{P} \mathbf{W} \mathbf{P}'$ and $\mathbf{P} \mathbf{V}_y \mathbf{P}'$ respectively containing the latent roots of \mathbf{W} and those of \mathbf{V}_y . It follows at once from (36.24) that these two sets of latent roots (which are all positive) coincide except for the first: if the roots of \mathbf{W} are λ_j , $j = 1, 2, \dots, n$, those of \mathbf{V}_y are

$\lambda_1 - n\theta^2 = \lambda_1^*$, say, and $\lambda_j, j > 1$. These latent roots are, of course, functions of the parameters σ_j^2 .

If s is the number of distinct roots of W , and s^* the number of distinct roots of V_y , we may have $s^* = s - 1$ when λ^* does not, coincide with some other root; or $s^* = s + 1$ if this situation is reversed; or $s^* = s$ if neither or both of λ_1, λ_1^* coincide with another root. Graybill and Hultquist (1961) show that $s \geq k + 2$; and that

$$s^* \leq 1 + \text{rank}(\mathbf{X}_0; \mathbf{X}_1; \dots; \mathbf{X}_k),$$

subject to a further condition.

36.11 Since $\mathbf{P}\mathbf{P}' = \mathbf{I}$, (36.22) is identical with (36.25)

$$Q = (\mathbf{P}\mathbf{y} - \mathbf{P}\mathbf{1}\theta)' (\mathbf{P}\mathbf{V}_y \mathbf{P}')^{-1} (\mathbf{P}\mathbf{y} - \mathbf{P}\mathbf{1}\theta).$$

We now partition \mathbf{P} into $\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_t$ where \mathbf{P}_1 is the first row as before, $\mathbf{P}_j (j > 1)$ is of order $(n_j \times n)$, and n_j is the multiplicity of the root λ_j when λ_1^* is ignored. Thus $t = s$ or $t = s + 1$. Using (36.24), (36.25) now becomes

$$Q = \begin{pmatrix} \mathbf{P}_1 \mathbf{y} - n^{\frac{1}{2}} \theta \\ \mathbf{P}_2 \mathbf{y} \\ \vdots \\ \mathbf{P}_t \mathbf{y} \end{pmatrix}' \begin{pmatrix} 1/\lambda_1^* & & 0 \\ & \mathbf{I}_2/\lambda_2 & \\ & & \ddots \\ 0 & & & \mathbf{I}_t/\lambda_t \end{pmatrix} \begin{pmatrix} \mathbf{P}_1 \mathbf{y} - n^{\frac{1}{2}} \theta \\ \mathbf{P}_2 \mathbf{y} \\ \vdots \\ \mathbf{P}_t \mathbf{y} \end{pmatrix}, \quad (36.26)$$

where \mathbf{I}_j is the identity matrix of order n_j .

We see at once from (36.26) that the t statistics

$$\mathbf{P}_1 \mathbf{y}, \quad \mathbf{y}' \mathbf{P}_j' \mathbf{P}_j \mathbf{y}, \quad j = 2, 3, \dots, t,$$

are sufficient for the $(k+2)$ parameters of the model, since no other statistic enters the Likelihood Function. It is intuitively obvious that they are *minimal* sufficient. The proof of this follows directly by the method of 23.18 if λ_1^* is not equal to any other of the λ_j ; if it is so equal, the proof is extended by Graybill and Hultquist (1961), who also use a result by Gautschi (1959) given in 36.16 below to show that if $t = s = k + 2$ (its smallest possible value—see 36.10) the set of minimal sufficient statistics is complete.

36.12 (36.26) shows directly that the statistic $\mathbf{P}_1 \mathbf{y}$ has a marginal univariate normal distribution with mean $n^{\frac{1}{2}} \theta$ and variance $\lambda_1^* = \mathbf{P}_1 \mathbf{V}_y \mathbf{P}_1'$. Further, each of the other $(t-1)$ components of the minimal sufficient set is a quadratic form

$$Q_j = \mathbf{y}' \lambda_j^{-1} \mathbf{P}_j' \mathbf{P}_j \mathbf{y}, \quad j = 2, 3, \dots, t,$$

in multinormal variables. Each matrix $\lambda_j^{-1} \mathbf{P}_j' \mathbf{P}_j \mathbf{V}_y$ is idempotent, since $\mathbf{P}_j \mathbf{V}_y \mathbf{P}_j' = \lambda_j \mathbf{I}_j$. Furthermore, since $\mathbf{P}_j \mathbf{V}_y \mathbf{P}_l' = \mathbf{0}, j \neq l$, the sum of the Q_j ,

$$Q^* = \sum_{j=2}^t \mathbf{y}' \lambda_j^{-1} \mathbf{P}_j' \mathbf{P}_j \mathbf{y}$$

also has the property that its matrix

$$\sum_{j=2}^t \lambda_j^{-1} \mathbf{P}_j' \mathbf{P}_j \mathbf{V}_y$$

is idempotent. Thus $Q^* = \sum_{j=2}^t Q_j$ is a decomposition of the type discussed in 35.7. The idempotencies just stated therefore imply the result that the $Q_j, j = 2, 3, \dots, t$, are distributed with degrees of freedom n_j in the χ^2 distribution, which is here central

because the non-central parameter is $(1\theta)' \lambda_j^{-1} \mathbf{P}_j' \mathbf{P}_j (1\theta) = 0$ by (36.24). Since the result above for $\mathbf{P}_1 \mathbf{y}$ is equivalent to the quadratic form $Q_1 = \mathbf{y}' \lambda_1^{-1} \mathbf{P}_1' \mathbf{P}_1 \mathbf{y}$ having a non-central χ^2 distribution with 1 d.fr. and non-central parameter $n\theta^2$, we see finally that the t quadratic forms Q_j have ranks adding to n , the rank of their sum Q at (36.26). Each is therefore independent of each of the others by 35.7.

Analysis of variance in Model II

36.13 We now observe an important distinction between Model I and Model II: in the latter, the SS $\mathbf{y}' \mathbf{y}$ cannot be decomposed into quadratic forms which are *themselves* independently distributed in the (central or non-central) chi-squared form, for $\mathbf{y}' \mathbf{y}$ itself is not so distributed, on account of the covariances between the observations—it is (36.22) which has that property instead. However, 36.11–12 show that the t quadratic forms Q_j are so distributed. The matrices of these forms are

$$\mathbf{P}_1' \mathbf{P}_1 / \lambda_1^*, \quad \mathbf{P}_j' \mathbf{P}_j / \lambda_j, \quad j = 2, 3, \dots, t,$$

and since $\sum_{j=1}^t \mathbf{P}_j' \mathbf{P}_j = \mathbf{P}' \mathbf{P} = \mathbf{I}$, we see that we may decompose $\mathbf{y}' \mathbf{y}$ into t quadratic forms which, when divided by the corresponding distinct latent roots of \mathbf{V}_y , are independent chi-squared variables. Moreover, the degrees of freedom are simply the multiplicities of the roots, and all the forms except Q_1 have central distributions.

36.14 Since

$$E(Q_j) = n_j, \quad j \geq 2,$$

we have at once

$$E(\mathbf{y}' \mathbf{P}_j' \mathbf{P}_j \mathbf{y} / n_j) = \lambda_j, \quad j \geq 2. \quad (36.27)$$

The latent roots may therefore be found as the expectations of the corresponding Mean Squares (MS) in the AV table. Since

$$E(\mathbf{P}_1 \mathbf{y}) = n^{\frac{1}{2}} \theta,$$

θ is estimated by the mean of all the observations y , as is obvious. We require to estimate the other $(k+1)$ parameters, the variances. If the λ_j are $(k+1)$ different functions of these parameters, (36.27) may be solved to give estimators of the parameters.

In the common case when the λ_j are all *linear* functions of the parameters, (36.27) is particularly easy to solve to give unbiased estimators of the σ_j^2 .

Graybill and Hultquist (1961) show that an AV in our extended sense exists with the latent roots all different functions of the parameters if and only if the commutativity condition (36.21) holds and $\mathbf{W} = E(\mathbf{y}\mathbf{y}')$ has $s = k+2$ (the minimum possible number—see 36.10) distinct latent roots. Under the multinormality assumption, the set of sufficient statistics is then complete (cf. 36.11) and the estimators are unique MV unbiased for their expectations, by 17.35. Graybill and Hultquist (1961) show that the MS in (36.27) remain MV unbiased quadratic estimators of their expectations under weaker conditions than multinormality.

36.15 We are now in a position to connect our study of Model II with the Model I AV of classifications investigated in Chapter 35. Any Model I AV table is a decomposition of the SS $\mathbf{y}' \mathbf{y}$ into component SS (of which one is for the general mean).

and the d.f. of the table (ranks of the quadratic forms) always add to n , the number of observations. If (36.21) is satisfied, the same decomposition (cf. 36.13) will hold in Model II, since these ranks remain additive, but it is now fundamentally (36.22) which is being decomposed as at (36.26), and it is the ratio of each SS (except the general mean) to the expectation of its MS which is distributed in the chi-squared form, which is now always central.

To every Model I AV (balanced, to satisfy (36.21)) there is therefore a corresponding balanced Model II AV.

Example 36.3 Balanced one-way classification

In the balanced one-way classification treated in Examples 36.1–2, the Model I AV table (35.24) will hold under Model II also, by 36.15. We require the expected values of the MS in the table, except that for the general mean.

It is immediately evident that the Residual SS, which we now call S_3 , satisfies the identity

$$S_3 = \sum_i \sum_q (y_{iq} - \bar{y}_{i.})^2 \equiv \sum_i \sum_q (\varepsilon_{iq} - \bar{\varepsilon}_{i.})^2,$$

and it has exactly the same distribution as in Model I, since it is free of θ and the u_i . The corresponding MS therefore has expected value $\sigma_\varepsilon^2 = \lambda_3$, say. The between-groups SS is denoted as at (35.21) by

$$S_2 = \sum_i n_i (y_{i.} - \bar{y}_{..})^2 \equiv n_i \sum_i \{(u_i + \bar{\varepsilon}_{i.}) - (u. + \bar{\varepsilon}_{..})\}^2;$$

n_i is a constant since we now have n observations in p groups(*) of equal size, $n_i = n/p$. Because the u 's and the ε 's are independent,

$$\text{var}(u_i + \bar{\varepsilon}_{i.}) = \sigma_1^2 + \frac{\sigma_\varepsilon^2}{n_i},$$

for $\varepsilon_{i.}$ is a mean of n_i error terms. Since $E(u_i + \bar{\varepsilon}_{i.}) = 0$, this shows directly that the variable

$$\sum_{i=1}^p \{(u_i + \bar{\varepsilon}_{i.}) - (u. + \bar{\varepsilon}_{..})\}^2 / (\sigma_1^2 + \sigma_\varepsilon^2 / n_i)$$

has the chi-squared form with $(p-1)$ d.f., since it is a standardized sum of squares about the sample mean. The expected value of S_2 , the between-groups SS, is therefore $(p-1)n_i(\sigma_1^2 + \sigma_\varepsilon^2/n_i)$ and that of the MS is $E\{S_2/(p-1)\} = \sigma_\varepsilon^2 + n_i\sigma_1^2 = \lambda_2$, say.

We thus have the two independently distributed chi-square variables

$$S_2/\lambda_2 = \sum_i n_i (y_{i.} - \bar{y}_{..})^2 / (\sigma_\varepsilon^2 + n_i\sigma_1^2), \quad S_3/\lambda_3 = \sum_i \sum_q (y_{iq} - \bar{y}_{i.})^2 / \sigma_\varepsilon^2,$$

whose ratio, after division by d.f. $((p-1)$ and $(n-p)$ respectively) will have the F -distribution. If $\sigma_1^2 = 0$, the denominators are identical, so that the same F -statistic (35.22) as in Example 35.1 may be used here to test $\sigma_1^2 = 0$, which is the Model II hypothesis of no difference between groups.

The same hypothesis in the two different models may thus be tested by the same statistic. But two points should be noted. First, we have as yet no assurance that

(*) We use p here, where k was used in Example 35.1, because k has another connotation in this chapter.

this is an *optimum* test in Model II, as we know it to be in Model I from general linear hypothesis theory. Second, although the test statistic is the same for both models, its distribution is the same *only* under the hypothesis of no difference between groups. Its power function must necessarily be different in the two models, since the alternatives are quite different in the two cases (cf. Exercise 36.1).

It will be seen by solving the expressions for $E(S_2)$ and $E(S_3)$ that

$$E\left\{\left(\frac{S_2}{p-1} - \frac{S_3}{n-p}\right) / n_i\right\} = \sigma_1^2.$$

This unbiased estimator of σ_1^2 can obviously be negative; it remains the MV unbiased estimator since it is a function of the complete sufficient statistics (\bar{y}, S_2, S_3) (cf. 36.11).

Example 36.4 Balanced two-way cross-classification

For the balanced two-way cross-classification, the Model I AV table ((35.65) and Exercise 35.1) will hold under Model II by 36.15, apart from its last column, which is specific to Model I. In the model (36.7), we now have $k = 3$. For convenience, we denote the three variances to be estimated by σ_R^2 , σ_C^2 and σ_{RC}^2 , indicating that they are respectively the variances of the Row, the Column and the Interaction (Row \times Column) random variables u . It is easy to see, as in Example 36.2, that the commutativity condition (36.21) holds; indeed, considerations of symmetry make this obvious.

Leaving aside the general mean, the four MS must have their expectations evaluated. As in Example 36.3, examination of the model, now written in an obvious notation

$$y_{ijp} = \theta + u_{i*} + u_{*j} + u_{ij} + \varepsilon_{ijp}, \quad (36.28)$$

$$(i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c; \quad p = 1, 2, \dots, m),$$

shows that the Residual SS, now denoted by $S_5 = \sum_i \sum_j \sum_p (y_{ijp} - y_{ij.})^2$, is identical with $\sum_i \sum_j \sum_p (\varepsilon_{ijp} - \varepsilon_{ij.})^2$ and has the same distribution as in Model I, so that its MS has expectation σ_ε^2 , and this will evidently be generally true for the balanced Model II.

The Rows SS is now written

$$S_2 = cm \sum_i (y_{i..} - y_{...})^2 \equiv cm \sum_i \{(u_{i*} + u_{i.} + \varepsilon_{i..}) - (u_{*} + u_{..} + \varepsilon_{...})\}^2,$$

and as in Example 36.3,

$$\text{var}(u_{i*} + u_{i.} + \varepsilon_{i..}) = \sigma_R^2 + \sigma_{RC}^2/c + \sigma_\varepsilon^2/(cm).$$

Thus the Rows MS has expectation

$$E\{S_2/(r-1)\} = \sigma_\varepsilon^2 + m\sigma_{RC}^2 + cm\sigma_R^2, \quad (36.29)$$

and by interchanging row and column symbols, we have similarly, for the Columns MS,

$$E\{S_3/(c-1)\} = \sigma_\varepsilon^2 + m\sigma_{RC}^2 + rm\sigma_C^2. \quad (36.30)$$

The Interactions SS is written

$$\begin{aligned} S_4 &= m \sum_i \sum_j (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2 \\ &\equiv m \sum_i \sum_j \{(u_{ij} + \varepsilon_{ij.}) - (u_{i.} + \varepsilon_{i..}) - (u_{.j} + \varepsilon_{.j.}) + (u_{..} + \varepsilon_{...})\}^2. \end{aligned} \quad (36.31)$$

The expression in braces on the right of (36.31) has zero mean, and the expectation of its square is therefore its variance. Now

$$\text{var}(u_{ij} - u_{i.} - u_{.j} + u_{..}) = \sigma_{RC}^2 \left(1 - \frac{1}{c}\right) \left(1 - \frac{1}{r}\right)$$

as may be seen by evaluating the four variances and six covariances required. Similarly

$$\text{var}(\varepsilon_{ij} - \varepsilon_{i.} - \varepsilon_{.j} + \varepsilon_{..}) = \frac{\sigma_{\varepsilon}^2}{m} \left(1 - \frac{1}{c}\right) \left(1 - \frac{1}{r}\right),$$

and thus the SS has expectation

$$E(S_4) = mrc \left(\sigma_{RC}^2 + \frac{\sigma_{\varepsilon}^2}{m} \right) \left(1 - \frac{1}{c}\right) \left(1 - \frac{1}{r}\right).$$

Dividing by the d.f. $(r-1)(c-1)$ and cancelling, this gives for the MS

$$E\{S_4 / [(r-1)(c-1)]\} = \sigma_{\varepsilon}^2 + m\sigma_{RC}^2. \quad (36.32)$$

We have thus obtained a column of expected MS to replace the final column of the table (35.65):

$$\left. \begin{array}{ll} \text{Rows:} & \lambda_2 = \sigma_{\varepsilon}^2 + m\sigma_{RC}^2 + cm\sigma_R^2 \\ \text{Columns:} & \lambda_3 = \sigma_{\varepsilon}^2 + m\sigma_{RC}^2 + rm\sigma_C^2 \\ \text{Interactions:} & \lambda_4 = \sigma_{\varepsilon}^2 + m\sigma_{RC}^2 \\ \text{Residual:} & \lambda_5 = \sigma_{\varepsilon}^2 \end{array} \right\} \quad (36.33)$$

(36.33) is immediately solvable to give unbiased estimators of the variance parameters. Apart from σ_{ε}^2 , any of these may be negative since they are differences, just as at the end of Example 36.3, and the remark made there applies here equally.

(36.33) tells us which functions of the parameters must be used as divisors to obtain central chi-squared variables from the SS in the AV Table. Examination of these shows that the happy coincidence of Model II and Model I tests found in the one-way classification of Example 36.3 no longer survives here. If we wish to test that either σ_R^2 (or σ_C^2) = 0, the expected MS of Rows (or Columns) will be equal to that for Interactions, not for Residual, and tests based on S_2/S_4 and S_3/S_4 are indicated. But we may test $\sigma_{RC}^2 = 0$ by the Interactions/Residual test of S_4/S_5 used in (35.65) for Model I. We find, for the first time, that we have to distinguish different choices of the divisor MS for the F -tests in an AV table.

Testing hypotheses in Model II AV

36.16 The remarks in Examples 36.3-4 draw our attention to the fact that we have given no theoretical justification so far for the use of F -tests in Model II AV. Even where the test coincides with that of Model I, its characteristics will be different and we cannot presume optimality; in any case, we have seen that new test statistics may be required. We must now consider the theory of Model II tests.

Remembering that we are specializing to AV situations where the $(t-1)$ quadratic forms in (36.26) are all Sums of Squares, we may write the multinormal distribution of the observations in the form

$$dG \propto \exp \left\{ -\frac{1}{2} \left[\frac{n(\bar{y} - \theta)^2}{\lambda_1^*} + \sum_{j=2}^t \frac{S_j}{\lambda_j} \right] \right\}. \quad (36.34)$$

Now we observe that λ_1^* , defined as $\lambda_1 - n\theta^2$, is not a function of θ at all, for by 36.12 it is equal to $\mathbf{P}_1 \mathbf{V}_y \mathbf{P}_1'$, i.e. n^{-1} times the sum of all the elements of \mathbf{V}_y , which are variances and covariances, and therefore free of θ . In every case we shall consider, $t = k+2$, the number of parameters to be estimated. Thus we have $(k+2)$ latent roots which are (in our case, always linear) functions of $(k+1)$ parameters σ_j^2 . We can eliminate this redundancy by expressing λ_1^* in terms of the other latent roots, so that (36.34) becomes

$$dG \propto \exp \left\{ -\frac{1}{2} \left[\frac{n\bar{y}^2 - 2n\bar{y}\theta}{p(\lambda_2, \dots, \lambda_t)} + \sum_{j=2}^{k+2} \frac{S_j}{\lambda_j} \right] \right\}. \quad (36.35)$$

Gautschi (1959) has shown the family of distributions

$$f(t, \tau) \propto \exp \left\{ \sum_{j=1}^r t_j \tau_j + t_1^2 g(\tau_2, \dots, \tau_r) \right\} \quad (36.36)$$

to be complete, a result not covered by (23.19), Vol. 2. We see at once that (36.35) is a case of (36.36) with

$$\begin{aligned} t_j &= S_j, \quad \tau_j = -1/(2\lambda_j), \quad j \geq 2; \\ t_1 &= \bar{y}, \quad \tau_1 = n\theta/p(\lambda_2, \dots, \lambda_t); \end{aligned}$$

and

$$g(\tau_2, \dots, \tau_r) = -\frac{1}{2}n/p(\lambda_2, \dots, \lambda_t).$$

Thus (36.35) is complete with this parametrization.

36.17 We may now make use of the results of Chapter 23. We are debarred from finding UMPU tests by the methods of 23.27–36, since (36.36) is more general than (23.73). However, we may find UMP similar tests directly by the method of 23.20.

In our applications to balanced Model II AV, the latent roots are linear forms in the parameters as exemplified in (36.33). The hypothesis which we wish to test (that one particular σ_j^2 , and that one alone, is zero) is always equivalent to testing that two particular latent roots are equal, as can be verified at (36.33) or in the simpler case of Example 36.3.

We first observe that $H_0 : \lambda_q = \lambda_p$; $q, p > 1$ leaves us with a set of $(k+1)$ complete sufficient statistics

$$T = \{\bar{y}, S_2, \dots, S_{p-1}, S_{p+1}, \dots, S_{q-1}, S_{q+1}, \dots, S_{k+2}, (S_p + S_q)\}.$$

Now we assume that $p(\lambda)$ in (36.35) is a function of λ_p or λ_q , but not of both. Following 23.20, we see that every similar region for H_0 consists of a fraction α of each contour of constant T , which we now hold fixed. We write

$$-\frac{1}{2} \left(\frac{S_p}{\lambda_p} + \frac{S_q}{\lambda_q} \right) = -\frac{1}{2} \left\{ \frac{S_p + S_q}{\lambda_q} + \frac{S_p(\lambda_q - \lambda_p)}{\lambda_p \lambda_q} \right\}. \quad (36.37)$$

For fixed T , use of the Neyman–Pearson lemma (22.6) on (36.35) with (36.37) inserted shows that the uniformly most powerful size- α critical region for testing H_0 against $H_1: \lambda_q > \lambda_p$ is given by

$$-\frac{1}{2} S_p (\lambda_q - \lambda_p) > c_\alpha(T),$$

so that the UMP critical region for any fixed T is given by small values of S_p , whatever the parameter values. Since $(S_p + S_q)$ is held fixed by T , this critical region is equivalent to large values of the ratio S_q/S_p . Finally, it is easy to see that this ratio is distributed free of the parameters when H_0 holds, and is therefore by Exercise 23.7 independent of the complete sufficient statistic T . Thus the test based on large values of S_q/S_p is unconditionally UMP similar.

Thus we have established that the UMP similar test of equality of two Expected Mean Squares in a (balanced) Model II AV table, against a one-sided alternative hypothesis, is the F -test of the ratio of the (potentially) larger to the smaller of these MS, large values leading to rejection of the hypothesis tested. Essentially this result was first given by Herbach (1959). Exercise 36.4 is to show that the tests are also UMPU. They are not LR tests in general (cf. Exercises 36.5-6).

Example 36.5

In Example 36.3, the Expected MS were:

$$\begin{array}{lcl} \text{Groups:} & \lambda_2 = \sigma_e^2 + n_i \sigma_1^2 & \\ \text{Residual:} & \lambda_3 = \sigma_e^2 & \end{array} \quad (36.38)$$

Provided that $\sigma_e^2 > 0$, as we always assume, the hypothesis that $\lambda_2 = \lambda_3$ is identical with $H_0: \sigma_1^2 = 0$, and $H_1: \lambda_2 > \lambda_3$ with $H_1: \sigma_1^2 > 0$. Thus the test given in Example 36.3 is UMP similar by 36.17. It turns out that in this case $\lambda_1^* = \lambda_2$ exactly, but this does not affect the test (see Exercises 36.2-3).

Example 36.6

In Example 36.4, (36.33) shows at once that if $\sigma_e^2 > 0$, $\sigma_R^2 = 0$ is equivalent to $\lambda_2 = \lambda_4$; $\sigma_C^2 = 0$ to $\lambda_3 = \lambda_4$, and $\sigma_{RC}^2 = 0$ to $\lambda_4 = \lambda_5$. The tests of these hypotheses indicated at the end of that Example are UMP similar by 36.17.

If we first test and accept the hypothesis that $\sigma_{RC}^2 = 0$ ($\lambda_4 = \lambda_5$), it is tempting to carry out the tests of $\sigma_R^2 = 0$ (or $\sigma_C^2 = 0$) by testing S_2 (or S_3) against the pooled SS ($S_4 + S_5$). Evidently the increase in the d.f. of the denominator of the variance ratio can bring an increase in the power of the test, but since the decision whether to pool S_4 with S_5 depends on the previous test, it may be wrongly taken when $\lambda_4 \neq \lambda_5$. As a result, control of the size of the overall test procedure becomes difficult. The numerically complicated theory, and recommendations for such pooling procedures, are treated by Bozivich *et al.* (1956) and Srivastava and Bozivich (1962).

Because the Interactions MS is the denominator for the tests of row-effects and of column-effects, we can make these tests, even when every cell frequency $m = 1$. This was not so in Model I (cf. Example 35.3) unless we were able to say that all interactions were zero. Here, only the test of $\sigma_{RC}^2 = 0$ is lost when all $n_{ij} = 1$ and the Residual SS is identically zero.

General balanced cross-classifications

36.18 The patterning of the expected MS given at (36.33) for the two-way cross-classification is suggestive for higher-order balanced cross-classifications. For the

three-way cross-classification, it will similarly be found that the MS have expectations given, in an obviously extended notation, by:

$$\begin{array}{lcl}
 & E(\text{MS}) & \\
 \text{Rows (R):} & \lambda_2 = \sigma_e^2 + m\sigma_{RCL}^2 + lm\sigma_{RC}^2 + cm\sigma_{RL}^2 + clm\sigma_R^2 & \\
 \text{Columns (C):} & \lambda_3 = \sigma_e^2 + m\sigma_{RCL}^2 + lm\sigma_{RC}^2 + rm\sigma_{CL}^2 + rlm\sigma_C^2 & \\
 \text{Layers (L):} & \lambda_4 = \sigma_e^2 + m\sigma_{RCL}^2 + cm\sigma_{RL}^2 + rm\sigma_{CL}^2 + rcm\sigma_L^2 & \\
 \text{First-order Interactions} & \left\{ \begin{array}{l} (R \times C): \lambda_5 = \sigma_e^2 + m\sigma_{RCL}^2 + lm\sigma_{RC}^2 \\ (R \times L): \lambda_6 = \sigma_e^2 + m\sigma_{RCL}^2 + cm\sigma_{RL}^2 \\ (C \times L): \lambda_7 = \sigma_e^2 + m\sigma_{RCL}^2 + rm\sigma_{CL}^2 \end{array} \right. & \\
 \text{Second-order Interactions} & (R \times C \times L): \lambda_8 = \sigma_e^2 + m\sigma_{RCL}^2 & \\
 \text{Residual:} & \lambda_9 = \sigma_e^2 &
 \end{array} \quad (36.39)$$

This corresponds to the model, generalizing (36.28),

$$y_{ijkp} = \theta + u_{i**} + u_{*j*} + u_{**k} + u_{ij*} + u_{i*k} + u_{*jk} + u_{ijk} + \varepsilon_{ijkp} \quad (36.40)$$

($i = 1, 2, \dots, r; j = 1, 2, \dots, c; k = 1, 2, \dots, l; p = 1, 2, \dots, m$)

with $\text{var}(u_{i**}) = \sigma_R^2, \dots, \text{var}(u_{ij*}) = \sigma_{RC}^2, \dots, \text{var}(u_{ijk}) = \sigma_{RCL}^2$.

The rule of formation of (36.39) (and (36.33) and (36.38)) is now clear. Any expected MS has σ_e^2 plus m times a linear function of the variances. This linear function contains every variance in the model which includes among its subscripts all the identifying letters of the MS. The coefficient of each such variance is the product of the upper limits of the suffixes in (36.40) corresponding to letters not included among the subscripts of the variance; if all letters are included, the coefficient is unity. Thus, e.g., considering the expectation for the MS for the $(R \times L)$ Interactions, the only variances containing both R and L among their subscripts are σ_{RL}^2 and σ_{RCL}^2 . σ_{RL}^2 omits only C from its subscripts, and the corresponding suffix in (36.40) is j , with upper limit c ; σ_{RCL}^2 includes all subscripts, and gets the coefficient unity. Thus we obtain $(c\sigma_{RL}^2 + \sigma_{RCL}^2)$, to be multiplied by m and added to σ_e^2 , as given in (36.39). More general rules of formation of expectations of MS, including this balanced Model II rule as a special case, are given by Cornfield and Tukey (1956) (see also Scheffé (1959)).

36.19 (36.39) reveals a new feature of the three-way cross-classification, which persists for all higher-order cases. In Examples 36.5–6, we saw that each hypothesis of interest in the one-way and two-way cases (that some variance was zero) was identical with the hypothesis of equality of two expected MS λ_j . In (36.39) it will be seen that this remains true so far as $\sigma_{RCL}^2, \sigma_{RC}^2, \sigma_{RL}^2$ and σ_{CL}^2 are concerned: these are respectively zero if and only if $\lambda_8 = \lambda_9, \lambda_5 = \lambda_8, \lambda_6 = \lambda_8$ or $\lambda_7 = \lambda_8$. Thus the second-order and all first-order interactions can be tested by UMP similar F -tests, using 36.17. The situation is different for the other variances σ_R^2, σ_C^2 and σ_L^2 .

Consider σ_R^2 , for example, contained in λ_2 . $H_0: \sigma_R^2 = 0$ cannot be expressed in terms of the equality of two λ_j . Instead we observe that $\lambda_2 + \lambda_8 \geq \lambda_5 + \lambda_6$ and that

H_0 is identical with

$$H_0: \lambda_2 + \lambda_8 = \lambda_5 + \lambda_6. \quad (36.41)$$

The theory of 36.17 is of no use now, and, so far as we know, no investigation of the optimum choice of test of (36.41) has been made. However, an approximate test may be based on the result of Exercise 36.7, and a similar approximation is clearly possible whenever we can express a hypothesis that a variance is zero equivalently as a hypothesis that a linear function of the λ_j is zero (see Exercise 36.8).

Hierarchical classifications

36.20 The general theory of 36.13–15 and 36.16–17 applies to balanced hierarchical, as well as cross-, classifications. In the balanced case, we may expeditiously make use of the fact that a hierarchical classification may be regarded as an incomplete cross-classification to obtain the additional column of expected MS required for the AV table.

Example 36.7 *Balanced two-way hierarchical classification*

Consider the model

$$y_{ijp} = \theta + u_i + u_{ij} + \varepsilon_{ijp} \quad (i = 1, 2, \dots, k; j = 1, 2, \dots, l; p = 1, 2, \dots, m) \quad (36.42)$$

corresponding to k groups at the first level of classification, (the same) l sub-groups within each of these, and m observations in each of the kl sub-groups. The AV table in Example 35.5 holds here, with the necessary changes in notation,* and we need only derive the expected value of each of its MS (except the general mean). We now note that (36.42) is a degenerate form of (36.28), in which we need only put $\sigma_0^2 \equiv 0$ and relabel u_{i*} as u_i . We also write σ_R^2 and σ_{RC}^2 as σ_1^2, σ_2^2 for our present purposes. With these changes, we find that λ_3 and λ_4 of (36.33) are equal, with a total multiplicity of $(l-1) + (k-1)(l-1) = k(l-1)$, the d.f. for sub-groups. (36.33) now gives:

$$\left. \begin{array}{ll} \text{Groups:} & \sigma_e^2 + m\sigma_2^2 + lm\sigma_1^2 \\ \text{Sub-groups:} & \sigma_e^2 + m\sigma_2^2 \\ \text{Residual:} & \sigma_e^2 \end{array} \right\} E(\text{MS}) \quad (36.43)$$

Evidently, UMP similar tests of $\sigma_1^2 = 0$ and of $\sigma_2^2 = 0$ are available from 36.17.

Example 36.8 *Balanced three-way hierarchical classification*

By specializing (36.40) in exactly the same way as in the previous Example, (36.39) becomes (the reader is left to do this as Exercise 36.9):

(*) Notice particularly that l in Example 35.5 corresponds to kl here, the total number of sub-groups.

	$E(MS)$	
Groups:	$\sigma_e^2 + m\sigma_3^2 + l_3 m\sigma_2^2 + l_2 l_3 m\sigma_1^2$	(36.44)
Sub-groups:	$\sigma_e^2 + m\sigma_3^2 + l_3 m\sigma_2^2$	
Sub-sub-groups:	$\sigma_e^2 + m\sigma_3^2$	
Residual:	σ_e^2	

Here, there are l_1 groups each containing l_2 sub-groups, each of which contains l_3 sub-sub-groups, and m observations in each of the latter, making $n = l_1 l_2 l_3 m$ in all.

Again there is no difficulty in obtaining tests by the method of 36.17—the problem which arose for higher-way cross-classifications in 36.19 was produced by the multiplicity of interactions, which do not enter into the present problem.

Scheffé (1959) gives an extended treatment of the three-way hierarchical case with possibly unequal numbers.

Power of tests, confidence intervals and negative estimates

36.21 By 36.17, a UMP similar F -test for the hypothesis that some variance is zero is equivalent to testing $H_0: \phi = 0$ in $\lambda_q = \lambda_p + \phi$ against $H_1: \phi > 0$. The ratio of $S_q/(\lambda_p + \phi)$ to S_p/λ_p always has a central F -distribution, so that

$$F = \frac{S_q}{S_p} \left(1 + \frac{\phi}{\lambda_p}\right)^{-1}. \quad (36.45)$$

This leads immediately to the power of the test of H_0 based on the statistic S_q/S_p , Exercise 36.1 being the simplest case.

36.22 Whereas in Model I we were led to consider multiple comparisons between the parameters (means) of the model, the natural next step in Model II is to consider confidence intervals for the variance parameters.

(36.45) leads immediately to confidence intervals for the parameter ϕ/λ_p , of which Exercise 36.11 is the simplest case. These intervals may cover (or even consist entirely of) negative values, as even that simplest case shows. This runs parallel to the possible negativeness of the point estimators of variances which we found in Examples 36.3–4. For practical purposes, a negative point estimate of a non-negative quantity is inadmissible, and is therefore usually replaced by the estimate zero (although this removes the unbiasedness of the estimator). Similarly, the negative portion of a confidence interval is usually replaced by the value zero.

Thompson (1962) gives an algorithm for obtaining non-negative estimates of variances which gives intuitively acceptable results in the one-way and two-way cross-classifications, but becomes complicated even in the three-way case.

Bulmer (1957)—see also Scheffé (1959)—obtains approximate confidence intervals for ϕ itself in (36.45), rather than the less useful intervals for ϕ/λ_p already mentioned (cf. Exercise 36.15).

There is no difficulty in constructing confidence intervals for the error variance σ_s^2 from the distribution of the Residual SS in every case; these are never negative.

The unbalanced case in Model II

36.23 Since Example 36.2, we have been confining ourselves to the balanced (equal frequencies) case. To exemplify the difficulties of the unbalanced case in Model II, and throw some light on their origin, we now return to the one-way classification, for which the model is given in Example 36.1.

36.24 We see that

$$\text{cov}(y_{iq}, y_{rs}) = 0 \quad \text{if } i \neq r \\ = \sigma_1^2 \quad \text{if } i = r,$$

so that the multinormality assumption implies that observations which are not in the same one of the p groups are independent, while those within the same group have covariance σ_1^2 ; all have zero means and variances $\sigma_1^2 + \sigma_\varepsilon^2$. (This is the simplest case of the general formula (36.11).)

To write down the exponent of the multinormal distribution, we must invert \mathbf{V}_y which, as we have just seen, contains only zeros apart from p square submatrices along its leading diagonals, of dimensions n_1, n_2, \dots, n_p (the numbers of observations in the p groups), each of which is of form

$$\mathbf{V}_{n_i} = \begin{pmatrix} \sigma_1^2 + \sigma_\varepsilon^2 & \sigma_\varepsilon^2 & \cdot & \cdot & \cdot & \sigma_\varepsilon^2 \\ \sigma_\varepsilon^2 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_\varepsilon^2 & \cdot & \cdot & \cdot & \cdot & \sigma_1^2 + \sigma_\varepsilon^2 \end{pmatrix}$$

with inverse

$$\mathbf{V}_{n_i}^{-1} = \{\sigma_\varepsilon^2(\sigma_\varepsilon^2 + n_i \sigma_1^2)\}^{-1} \begin{pmatrix} \sigma_\varepsilon^2 + (n_i - 1)\sigma_1^2 & -\sigma_1^2 & \cdot & \cdot & \cdot & -\sigma_1^2 \\ -\sigma_1^2 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ -\sigma_1^2 & \cdot & \cdot & \cdot & \cdot & \sigma_\varepsilon^2 + (n_i - 1)\sigma_1^2 \end{pmatrix}$$

which may be verified by multiplication. Thus \mathbf{V}_y^{-1} contains the $\mathbf{V}_{n_i}^{-1}$ along its leading diagonal, and zeros otherwise.

36.25 The logarithm of the LF for the one-way classification is, from 36.24,

$$\log L = c(\sigma_1^2, \sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2} \left\{ \sum_{i=1}^p \frac{\sigma_\varepsilon^2 + (n_i - 1)\sigma_1^2}{\sigma_\varepsilon^2 + n_i \sigma_1^2} \sum_{q=1}^{n_i} (y_{iq} - \theta)^2 \right. \\ \left. - \sum_{i=1}^p \frac{\sigma_1^2}{\sigma_\varepsilon^2 + n_i \sigma_1^2} \sum_{q \neq l} \sum (y_{iq} - \theta)(y_{il} - \theta) \right\} \\ = c(\sigma_1^2, \sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2} \left\{ \sum_i \sum_q (y_{iq} - \theta)^2 - \sum_i \frac{\sigma_1^2}{\sigma_\varepsilon^2 + n_i \sigma_1^2} [\sum_q (y_{iq} - \theta)]^2 \right\} \\ = c(\sigma_1^2, \sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2} \left\{ \sum_i \sum_q (y_{iq} - y_{..})^2 + n(y_{..} - \theta)^2 - \sum_i \frac{\sigma_1^2}{\sigma_\varepsilon^2 + n_i \sigma_1^2} [n_i(y_{i.} - \theta)]^2 \right\}. \quad (36.46)$$

Since $y_{..} = \sum_{i=1}^p n_i y_{i.}/n$, it is obvious from (36.46) that the $(p+1)$ statistics $\{\sum_i \sum_q (y_{iq} - y_{..})^2; y_{i.}, i = 1, 2, \dots, p\}$ are sufficient for the three parameters of the problem. If the n_i are all equal, the last summation on the right of (36.46) may be written

$$-\frac{n_i \sigma_1^2}{\sigma_e^2 + n_i \sigma_1^2} \sum_i n_i (y_{i.} - \theta)^2 \equiv -\frac{n_i \sigma_1^2}{\sigma_e^2 + n_i \sigma_1^2} \left\{ \sum_i n_i (y_{i.} - y_{..})^2 + n(y_{..} - \theta)^2 \right\}, \quad (36.47)$$

and (36.47) inserted into (36.46) shows that we then have the minimal sufficient set of three statistics $\{\sum_i \sum_q (y_{iq} - y_{..})^2, y_{..}, \sum_i n_i (y_{i.} - y_{..})^2\}$, in accordance with the general result of 36.11 for the balanced case. These are, of course, essentially the quantities entering into the AV table discussed in Example 36.3.

However, if the n_i are not all equal, (36.47) does not hold, and the minimal sufficient statistic has more than three components (cf. Exercise 36.12).

36.26 The AV is severely affected by the lack of balance. The "between groups" SS $S_2 = \sum_{i=1}^p n_i (y_{i.} - y_{..})^2$ is no longer distributed as a multiple of a chi-squared variable, for it is a weighted sum of squares about their mean of normal variables with zero means but unequal variances. However, the distribution of the Residual SS S_3 is unchanged, so that

$$E\{S_3/(n-p)\} = \sigma_e^2 \quad (36.48)$$

as before.

One can still estimate σ_1^2 from the AV table, but this is no longer a unique optimum procedure, as it was in the balanced case of Example 36.3. We saw in 36.25 that S_3 and the p group means are always sufficient statistics. Consider the function of the group means

$$S(m_i) = \sum_{i=1}^p m_i y_{i.}^2 - \left(\sum_{i=1}^p m_i \right) y_{..}^2,$$

where the m_i are constants. Since, from Example 36.3,

$$\text{var } y_{i.} = \sigma_1^2 + \sigma_e^2/n_i, \quad (36.49)$$

we see that

$$E\{S(m_i)\} = \sigma_e^2 \sum_i m_i \left(\frac{1}{n_i} - \frac{1}{n} \right) + \sigma_1^2 \left(1 - \frac{\sum_i n_i^2}{n^2} \right) \sum_i m_i. \quad (36.50)$$

(36.50) can be solved with (36.48) to give an unbiased estimator of σ_1^2 , whatever the m_i used, and we thus obtain a multiplicity of unbiased estimators of σ_1^2 (except in the balanced case (cf. Exercise 36.13)). The "natural" choice $m_i = n_i$, which reduces $S(m_i)$ to S_2 , is convenient but has nothing else in general to recommend it.

This lack of uniqueness demonstrates that (as usual when the dimension of the vector of sufficient statistics exceeds the number of parameters) we have lost the completeness of the sufficient statistic in the unbalanced case.

36.27 Tukey (1956-7) considered the problem of optimum estimation in the unbalanced one-way classification, with complicated results which should be consulted. Searle (1958) and Low (1964) investigated the unbalanced two-way cross-classification, where again the SS in the AV table are no longer chi-squared multiples. Henderson (1953) considered several methods of unbiased estimation in the general unbalanced case.

Generalization of AV models: discussion

36.28 Both the LS general linear Model I of Chapter 35 and Model II of the present chapter are "extreme" models, in the sense that *all* the elements of the vector θ in (19.8) are constants (parameters) while *all* the elements of \mathbf{u} in (36.1) are random variables, uncorrelated with each other and with the error-vector ϵ . In practice, we may meet situations where a mixture of the two models is apposite, i.e. where

$$\mathbf{y} = 1\theta_0 + \mathbf{X}_1\theta + \mathbf{X}_2\mathbf{u} + \epsilon, \quad (36.51)$$

where θ_0 is now used for the general mean to distinguish it from the parameter-vector θ of constants.

36.29 If we confine ourselves from the outset to discussion of AV situations (which we did not do until Chapter 35 for Model I and until **36.13** for Model II), we can easily see how "mixed" models of the form (36.51) may arise. Consider for illustration a two-way cross-classification.

Suppose that an experiment is to investigate the breaking strengths of five different types of paper when wet, and that three different levels of moisture content are to be used with each type of paper. This will give a table with five rows and three columns for the results of the tests. The five types of paper have been selected for the experiment because they are of intrinsic interest—we want to know how these papers behave. It is therefore reasonable to regard their breaking strengths as constants (parameters) subject to the usual experimental errors of determination. So far as the row-classification is concerned, Model I is quite appropriate.

The situation may be different for the column-classification. The three levels of moisture-content will probably have been chosen as convenient levels to represent "high," "medium" and "low" content, there being nothing sacrosanct about the precise levels used. In this sense, the three levels used will have been chosen from a population of potential levels of moisture-content in some (not necessarily probabilistic) way. Thus the column-effects will have some kind of distribution, quite apart from experimental error. So far as the column-classification is concerned, therefore, Model II (without the normality assumption) is a more reasonable idealization of the experiment, though by no means a perfect one. We are therefore led in the first instance to represent this experiment by a model of the form (36.51), with θ standing for row-effects and \mathbf{u} for column-effects.

36.30 We may, further, consider the interactions of rows and columns in the experiment of **36.29**, i.e., as in **35.18**, allow for the possibility that the five types of paper have different relative patterns of breaking strengths at the different levels of moisture-content. Since the column-effects are themselves idealized as random variables, it seems logically necessary that their interactions with the rows must also be treated in this way, rather than as constants. We easily achieve this by allowing \mathbf{u} in (36.51) to have further components to represent the interactions. However, this leads us immediately to consider a point which also applies to Model II itself: if we allow that column-effects and row-column interactions are random variables, what justification can there be for assuming that they are uncorrelated, especially when we

recall that the introduction of the normality assumption into the model then implies independence?

On grounds of realism, independence of random effects and interactions can hardly be allowed, and Model II, while retaining its mathematical interest, seems unlikely to do full justice to many practical situations involving interactions.

36.31 The fact that Model II AV is identified with the generally unacceptable assumption of *independent interactions*, as we shall now call it, stimulated the development of other more general AV models which could be freed from this assumption. We shall see that models with *tied interactions*, i.e. interactions correlated with the corresponding main effects, do indeed lead to a different AV procedure from that of either Model I or Model II. Recent expositions, with some historical detail, are given by Plackett (1960) and Scheffé (1956b).

We now examine AV models developed by Cornfield and Tukey (1956) (following earlier unpublished work by these authors), Scheffé (1956a), and by Wilk and Kempthorne (1955-6). We confine our discussion to the two-way cross-classification, which exhibits all the important features of the models.

A general model

36.32 Suppose that we have a population (discrete or continuous) of possible levels for the row-classification, from which r levels are selected for use in an experiment, and call this population P_R . Similarly, suppose that there is a population P_C of possible levels for the column-classification, from which the c levels used are selected. The selection process for rows is assumed independent of that for columns. If row-level i and column-level j are selected, n_{ij} observations are to be made on this combination. The p th observation in the i th row and j th column will be written y_{ijp} as before, and we let

$$y_{ijp} = \mu_{ijp} + \varepsilon_{ijp}, \quad (36.52)$$

where $E(\varepsilon_{ijp}) = 0$. We leave aside for the moment the detailed consideration of how the n_{ij} required experimental units are to be allocated to the (i, j) th selected row-column combination (see 36.36 and 36.39 below), but we let N_{ij} denote the number of experimental units which could (by the structure of the experiment) so be allocated, and define μ_{ij*} to be the average value of μ_{ijp} calculated over its N_{ij} possible values. Further, μ_{i**} and μ_{*j*} are averages of μ_{ij*} over all the levels in P_C and all the levels in P_R respectively, while μ_{***} is an average of μ_{ij*} over both P_R and P_C . No assumption of normality or homoscedasticity is made.

36.33 The general mean, row-effects, column-effects and row-column interactions are now respectively defined, by analogy with (35.31), as

$$\left. \begin{aligned} \theta_{**} &= \mu_{***} \\ \theta_{i*} &= \mu_{i**} - \mu_{***} \\ \theta_{*j} &= \mu_{*j*} - \mu_{***} \\ \theta_{ij} &= \mu_{ij*} - \mu_{i**} - \mu_{*j*} + \mu_{***} \\ &= \mu_{ij*} - \theta_{i*} - \theta_{*j} - \theta_{**} \end{aligned} \right\} \quad (36.53)$$

Although only r values of i and c values of j will be observed, all our definitions of the θ 's are made in terms of the μ 's, which are functions of the *populations* of levels, rather than only those actually appearing in the experiment as it is carried out. Further, the θ_{ij} are *tied interactions*, related to the population effects θ_{i*} and θ_{*j} through the last line of (36.53).

From (36.53) we may now define the components of variation. If P_R has R members and P_C has C members, we write

$$\left. \begin{aligned} \sigma_R^2 &= \sum_{i=1}^R \theta_{i*}^2 / (R-1), \\ \sigma_C^2 &= \sum_{j=1}^C \theta_{*j}^2 / (C-1), \\ \sigma_{RC}^2 &= \sum_{i=1}^R \sum_{j=1}^C \theta_{ij}^2 / \{(R-1)(C-1)\}, \end{aligned} \right\} \quad (36.54)$$

where R or C or both may be allowed to tend to infinity, e.g. in the continuous case. The residual variance is defined by

$$\sigma_e^2 = \frac{1}{RC} \sum_{i=1}^R \sum_{j=1}^C (N_{ij}-1)^{-1} \sum_{p=1}^{N_{ij}} (\mu_{ijp} - \mu_{ij*})^2. \quad (36.55)$$

36.34 We now specialize the general model in various ways.

Case 1

P_R has $R = r$ members only, and P_C has $C = c$ members only, and no sampling of row- and column-levels takes place; $N_{ij} \rightarrow \infty$ for all i, j .

This is the basis for Model I AV in Chapter 35. (36.53) becomes identical with (35.31).

Case 2

Both P_R and P_C are continuous populations, so that R and $C \rightarrow \infty$, as does N_{ij} for all i, j .

This is the basis for Model II AV in this chapter, although we have here made no assumption of normality or homoscedasticity. The last equation of (36.53) gives

$$\mu_{ij*} = \theta_{**} + \theta_{i*} + \theta_{*j} + \theta_{ij}, \quad (36.56)$$

and all the terms on the right except θ_{**} are random variables. The interactions θ_{ij} are now effectively made independent of the θ_{i*} and θ_{*j} by the infiniteness of the population from which they are sampled. In fact, (36.56) is equivalent, apart from changes of notation, to (36.28) (written for y_{ij} instead of the individual observations y_{ijp}).

Case 3

The observed r row-levels are selected at random without replacement from the R levels in P_R , and the observed c column-levels are, independently of the row selections, similarly selected from the C levels in P_C .

This is a model in which both classifications have random effects, as in Model II, but with tied interactions. It reduces to Case 2 when R, C and $N_{ij} \rightarrow \infty$.

It should be noted that even when $r = R$, $c = C$ and $N_{ij} \rightarrow \infty$, Case 3 is not identical with Case 1 above, since there was no sampling at all in Case 1, whereas here the sampling process determines the order in which the r rows and c columns are labelled $1, 2, \dots, r$ and $1, 2, \dots, c$ respectively. The sampling process effects a randomization of the rows and (independently) of the columns, of the $r \times c$ table. We shall call this Case 3(1).

Mixed models

36.35 Case 4

In Case 3 above, suppose that $R = r$, so that there is only a permutation of rows, which are otherwise fixed, but that, as in Case 2, C and $N_{ij} \rightarrow \infty$.

This is a *mixed model*, with row-levels fixed apart from permutations but the column-levels selected from an infinite population. The column-vector $\{\mu_{ij*}\}$, with r components obtained by giving i the values $1, 2, \dots, r$, is a random vector because of the selection of columns. For different j , the $\{\mu_{ij*}\}$ are assumed mutually independent with the same multivariate distribution. However, this model is not as general as might at first appear, for the row-permutation process implies that the variances of the elements of $\{\mu_{ij*}\}$ are all equal, and similarly that their $\frac{1}{2}r(r-1)$ covariances are equal. This condition of complete symmetry is clearly not always desirable in practice. We therefore consider a further generalization, due to Scheffé (1956a), which we shall call Case 5.

Case 5

$R = r$ as in Case 1, with no sampling, $C \rightarrow \infty$, and as in Case 4, the $\{\mu_{ij*}\}$ are identically distributed random vectors. The dispersion matrix of their r components no longer has complete symmetry necessarily imposed upon it by a row-permutation process as in Case 4. This is therefore a more general mixed model than that of Case 4.

Imhof (1960) generalizes Case 5 to the balanced three-way classification with two random (and one fixed) classification variables.

36.36 Cornfield and Tukey (1956) give a detailed treatment of Case 3 of **36.34** (of which Cases 2 and 4 are specializations) not only for the two-way cross-classification to which we are confining ourselves, but for general balanced classifications. In all cases, they assume that for each selected row-column combination the n_{ij} observations are made upon experimental units selected at random without replacement from a distinct population of N_{ij} experimental units, and that these rc separate populations are all sampled independently of the row- and column-levels selections already discussed. For the balanced two-way cross-classification with all $n_{ij} = m$ and also all $N_{ij} = N$, their results for the expected MS in the AV table are given in the first two columns of (36.57). The remaining three columns specialize these results as indicated on the following page.

THE ADVANCED THEORY OF STATISTICS

Expected values of MS in AV table

		Special cases of Case 3		Case 4
Case 3		Case 3(1)	Case 2	
Rows	$\sigma_e^2 \left(1 - \frac{m}{N}\right) + m\sigma_{rc}^2 \left(1 - \frac{c}{C}\right) + cm\sigma_R^2$	$\sigma_e^2 + cm\sigma_R^2$	$\sigma_e^2 + m\sigma_{Rc}^2 + cm\sigma_R^2$	$\sigma_e^2 + m\sigma_{Rc}^2 + cm\sigma_R^2$
Columns	$\sigma_e^2 \left(1 - \frac{m}{N}\right) + m\sigma_{rc}^2 \left(1 - \frac{r}{R}\right) + rm\sigma_C^2$	$\sigma_e^2 + rm\sigma_C^2$	$\sigma_e^2 + m\sigma_{Rc}^2 + rm\sigma_C^2$	$\sigma_e^2 + rm\sigma_C^2$
Interactions	$\sigma_e^2 \left(1 - \frac{m}{N}\right) + m\sigma_{Rc}^2$	$\sigma_e^2 + m\sigma_{Rc}^2$	$\sigma_e^2 + m\sigma_{Rc}^2$	$\sigma_e^2 + m\sigma_{Rc}^2$
Residual	σ_e^2	σ_e^2	σ_e^2	σ_e^2

(36.57)

It will be seen that the entries in the Case 2 column are identical with (36.33) in Example 36.4, i.e. with the Model II results, as stated under Case 2 in 36.34.

On the other hand, despite the distinction made between Case 3(1) and Case 1 in 36.34, the entries in the Case 3(1) column in (36.57) are identical with the Model I results in Example 35.2 at (35.65), when we recall that a non-central chi-squared variate has mean value equal to the sum of its non-central parameter and its d.f. and use the definitions (36.54) with $R = r$, $C = c$. The randomization of rows and columns in Case 3(1) does not affect the expected MS because the SS are symmetric in row-levels and in column-levels.

36.37 The entries under Case 4 in (36.57) are new to us, and we see that the expected MS for Rows (which are fixed apart from permutation) is identical with its value in Model II (where row-effects are random), while the expected MS for Columns (which are random here) is identical with its value in Model I (where column-effects are fixed). In fact, the expected MS for Rows is determined by the sampling in the columns classification, and vice versa. The Rows MS is concerned with differences between row-levels, and the whole population of these is observed, but in association with only a sample of column-levels; the population interaction between row- and column-levels is therefore relevant, and σ_{Rc}^2 appears in the expected MS. The observed Columns MS does not depend on that population interaction, and σ_{Rc}^2 does not enter into the expectation of this MS. The phenomenon is general in more complex classi-

fications—the sampling of levels of the other classifications than the one under consideration determines the structure of the latter's expected MS.

36.38 The expected MS in Case 5 are the same as those given for Case 4 in (36.57). As in the Case 3(1) discussion at the end of 36.36, this is because the SS are symmetric functions of the rows. However, an important distinction between Cases 4 and 5 arises as soon as we introduce the multinormality assumption upon the $\{\mu_{ij*}\}$. It then appears (see Scheffé (1956a, 1959) for a detailed discussion) that the mixed model of Case 5 does not retain the simplicity of Model II, where we found (cf. 36.15–17) that the ratio of each SS in the AV table to its expected MS has a central chi-squared distribution, and as a consequence were able to obtain UMP similar F -tests by testing each MS against another with the same expectation on the hypothesis. It remains true in Case 5, as the last column of (36.57) suggests, that (if $m > 1$) $\sigma_{RC}^2 = 0$ may be tested by an F -test on Interactions MS/Residual MS, and $\sigma_C^2 = 0$ may be tested similarly. But the statistic Rows MS/Residual MS does not in general have an F -distribution, even though its numerator and denominator are independent with the same expectations when $\sigma_R^2 = 0$. Scheffé (1956a, 1959) gives an alternative test statistic distributed in the Hotelling's T^2 form to be discussed generally in 41.15–17. In Case 4, however, the statistic Rows MS/Residual MS remains distributed in the variance-ratio form—the symmetry saves the distribution.

S. N. Roy and Cobb (1960) consider the mixed model (no interactions) with normal errors and one or more random effects which are non-normally distributed.

We mention briefly that some sequential AV procedures have been developed by D. R. Cox (1952), Johnson (1953–4) and Ghosh (1964), whose paper should be consulted for further references in this field.

Allocation of experimental units: randomization

36.39 Despite the complications into which the proliferation of AV models has led us, we have not even begun to consider an important source of variability in most experimental data.

In 36.32 we left aside the question of the allocation of experimental units to the various row-column combinations to be used, and we saw in 36.36 that the general model there designated as Case 3 assumed that the n_{ij} units allocated to a selected row-column combination come from a distinct population of N_{ij} units, so that there are rc populations of experimental units. This is an extreme situation—the populations of units do not overlap at all.

At the other extreme is the situation where all the experimental units to be used (e.g. rcm in the balanced two-way classification) are selected at random without replacement from a single population of N units. Here $N_{ij} = N$ for all i, j , and there is complete overlap, so to speak. This method of allocation is called *complete randomization*, and any experiment employing it is a *completely randomized experiment*.

There are also obviously intermediate situations of partial overlap, where groups of row-column combinations share the same population of experimental units, and there are more than one but less than rc such populations. These would still be called *randomized* allocations (though not “completely randomized”). For example, each

row (or each column) of the experiment may have its own population of experimental units—this is the case when the allocation is in *randomized blocks*, to which we shall return in Chapter 38.

36.40 The reader will perhaps wonder why, in 36.39, the term “randomized” is denied only to the separate-populations method of allocation, for, after all, each of these populations is sampled at random. Like most confusing usage, this can be understood from the early history of the subject. The early formulations of AV did not explicitly allow for sampling of the separate-populations kind—(36.57) indeed shows that so long as each population is large, its size exercises little influence on the Case 3 AV table. On the other hand, questions of efficiency in experiment design (to which we shall turn in Chapter 38) forced early consideration of randomized blocks and similar methods of allocating experimental units, and here the explicit and essential randomization procedure became eponymous.

36.41 Randomization of experimental units should clearly be taken into account in the analysis of the data, but this leads to considerable complications, and we need some new definitions.

Explicitly, we wish to allow for the possibility that μ_{ijp} , defined at (36.52) as the mean value of the p th observation on the (i,j) th row-column combination, may itself depend upon the characteristics of the particular experimental unit upon which the p th observation is made, as well as upon (i,j) . Consider any group of row-column combinations which share the same population of experimental units, as discussed in 36.39, and suppose there are m members of this group, and that their population contains M experimental units, $M \geq m$. We now make a formal two-way classification of group-members against units. We emphasize that (i,j) is now being treated as a single classification by bracketing these suffixes on the right of the identity

$$\mu_{ijp} \equiv \mu_{(*)} + \{\mu_{(ij)*} - \mu_{(*)}\} + \{\mu_{(*)p} - \mu_{(*)}\} + \{\mu_{(ij)p} - \mu_{(ij)*} - \mu_{(*)p} + \mu_{(*)}\}, \quad (36.58)$$

$$\equiv \mu_{(ij)*} + \{\mu_{(*)} - \mu_{(*)}\} + \{\mu_{(ij)p} - \mu_{(ij)*} - \mu_{(*)p} + \mu_{(*)}\}, \quad (36.59)$$

where asterisks denote averaging as before.

(36.58) resolves μ_{ijp} into a “general mean,” two “main effects,” and an “interaction.” If both terms in braces in (36.59) are identically zero, the allocation of experimental units to row-column combinations is irrelevant to μ_{ijp} , for then it equals its average over all units. The first term in braces in (36.59) is called the *unit error* of the experimental unit concerned. The second term in braces there will be called the *interactive error* for the experimental unit and row-column combination concerned. (More usually it is called the unit-treatment interaction.)

36.42 When the resolution of μ_{ijp} into three components at (36.59) is superposed on the underlying two-way classification scheme set out in 36.32–3, the model becomes complicated. The mere fact that the interactive error term in (36.59) carries suffixes i, j and p , as does also ε_{ijp} in (36.52) (which we now call the *technical error*, to distinguish it from the unit and interactive errors defined above), leads us to expect difficulties in estimation of the various components of the model.

It is worth emphasizing that the technical error alone is an error term in the usual

sense, arising from inaccurate measurement or observation. The unit and interactive errors arise purely from the allocation of experimental units to the row-column combinations.

36.43 Wilk and Kempthorne (1955-6) discuss the one-, two- and three-way cross-classification in Case 3 of 36.34, including the unbalanced case, when there is complete randomization of experimental units. For the case of proportional frequencies, an orthogonal AV is always possible (cf. 35.21-2). For general (non-proportional) frequencies, a non-orthogonal AV using unweighted means of levels is used (cf. 35.31). The difficulties anticipated at the end of 36.42 duly arise—only certain functions of the parameters can be estimated. Moreover, as Plackett (1960) remarks, it is hardly natural to regard *unequal* frequencies n_{ij} as fixed numbers when row- and column-levels are being sampled; however, the addition of yet another sampling process to determine the n_{ij} (which might also be correlated with the observed values of y) would complicate the analysis even further.

36.44 We shall defer further discussion of randomization models until Chapter 38, where we shall examine their *rationale*. We have examined their effects on AV procedures sufficiently closely for our present purpose, and we conclude this chapter with some discussion of the implications of its contents.

The choice of an AV model

36.45 The plethora of models now available for AV presents the applied statistician with a problem which, in less acute forms, arises in the use of statistical techniques generally. Evidently, careful analysis of the known facts concerning the origins of the observations must be undertaken before a model can be chosen which reasonably represents the real-life situation; and where there is little such knowledge, a good deal of guesswork may be necessary. In this respect, the statistician is experiencing a situation familiar in almost every field of applied science.

It is worth pointing out that the varieties of AV discussed in 36.32-8 differ in their assumptions about the methods of selection of the levels of the factors being analysed, and not in any assumptions about the real nature of these factors or of the variables underlying them. Provided that the data arise from a designed experiment, the assumptions are concerned with the behaviour of the experimentalist rather than that of his material. On the other hand, the complications of 36.39-43 are essentially concerned with the nature of the material being experimented on. It must always be a matter for the experimenter to judge whether his experimental units differ enough to make these added complications in the analysis worth while—in social, agricultural and biological work they sometimes do, and in physical and industrial experimentation they often do not. Our present point is that, even when the observations arise by deliberate design, there is ample scope for intuitive skill in such judgements. *A fortiori*, when the observations are not the result of a designed experiment, the validity of the chosen analysis will depend on the insight of the statistician.

36.46 It will be clear, then, that AV, like other statistical techniques, is not a mill which will grind out results automatically without care or forethought on the part of

the statistician. It is, rather, an assortment of delicate instruments which can be brought into use when appropriate. It requires skill, as well as hard work, in use. Elaborate techniques need not be (though they sometimes have been) applied to prove something which was almost obvious to inspection from the start—the statistician must never lose sight of the need to scrutinize his material. Equally, inappropriate analyses have often been made. AV has no monopoly of the misapplications of statistics, but the multiplicity of models now available makes it particularly vulnerable to the errors of the single-minded enthusiast.

36.47 In the next chapter, the last of our three concerned with AV techniques, we first investigate the problem of transforming the data so that AV may be used. This leads to a discussion of the robustness of AV and of distribution-free methods in this field. Finally, we shall there consider the difficulties produced by incomplete data.

EXERCISES

36.1 In Example 36.3, show that the power function of the size- α test of $H_0: \sigma_1^2 = 0$ against $H_1: \sigma_1^2 > 0$ is

$$P(\sigma_1^2) = 1 - G \left\{ \left(1 + n_i \frac{\sigma_1^2}{\sigma_e^2} \right)^{-1} G_\alpha \right\},$$

where G is the distribution function of the central F -distribution with appropriate degrees of freedom and G_α the value exceeded with probability α , so that $G\{G_\alpha\} = 1 - \alpha$. Show that $P(\sigma_1^2)$ is monotonic in its argument, that the test is unbiased, and that it is consistent as group size n_i increases to infinity, but not if the number of groups p alone $\rightarrow \infty$.

36.2 In 36.17, show that since \bar{y} is a component of T , the term $\exp \{-\frac{1}{2}n\bar{y}^2/p(\lambda)\}$ in (36.35) will not affect the UMP similar test of H_0 against H_1 even if $p(\lambda)$ is a constant multiple of λ_p or λ_q .

36.3 In (36.26), show that the sum of all n latent roots, $\lambda_1^* + \sum_{j=2}^t \lambda_j n_j$, equals the sum of the variances of the n observations. Hence show in Example 36.5 that $\lambda_1^* = \lambda_2$.

36.4 Show that the UMP similar one-sided F -tests of 36.17 are unbiased, and are thus UMPU size- α tests.

36.5 For the balanced one-way classification in Examples 36.3 and 36.5, show that the ML estimators of λ_2 and λ_3 are

$$\hat{\lambda}_2 = S_2/p, \quad \hat{\lambda}_3 = S_3/\{p(n_i - 1)\} \quad \text{if } \hat{\lambda}_2 \geq \hat{\lambda}_3$$

but become

$$\hat{\lambda}_2 = \hat{\lambda}_3 = \{p\hat{\lambda}_2 + p(n_i - 1)\hat{\lambda}_3\}/n \quad \text{if } \hat{\lambda}_2 < \hat{\lambda}_3.$$

Hence show that the LR test statistic l for $H_0: \lambda_2 = \lambda_3$ is given by

$$l^2 = \frac{\hat{\lambda}_2^p \hat{\lambda}_3^{p(n_i - 1)}}{\hat{\lambda}_2^n} \quad \text{if } \hat{\lambda}_2 \geq \hat{\lambda}_3$$

$$= 1 \quad \text{if } \hat{\lambda}_2 < \hat{\lambda}_3,$$

so that $l = 1$ whenever the test statistic $F < \frac{p}{p-1}$, and the LR test is not equivalent to the F -test.

Show further that for $F \geq \frac{p}{p-1}$, l is a monotone decreasing function of F . Verify that the critical values $F_{\alpha}\{p-1, p(n_i-1)\}$ always exceed $\frac{p}{p-1}$ for $\alpha \leq 0.25$, so that for all practical purposes the LR test is equivalent to the UMP similar F -test. (Herbach, 1959)

36.6 For the balanced two-way cross-classification of Examples 36.4 and 36.6, show that the LR test of $H_0: \sigma_R^2 = 0$ is a function of S_2, S_3 and S_4 , whereas the UMP similar F -test is a function of S_2/S_4 only, and thus that the tests are not equivalent. (Herbach, 1959)

36.7 If S_j/λ_j are independently distributed χ^2 variables with ν_j d.fr., and $\lambda = \sum_j c_j \lambda_j$, show that $S = \sum_j (c_j S_j/\nu_j)$ has its first two moments identical with those it would have if $\nu S/\lambda$ were also a χ^2 variable with d.fr.

$$\nu = \lambda^2 / \sum_j (c_j^2 \lambda_j^2 / \nu_j),$$

estimated by

$$\nu^* = S^2 / \sum_j (c_j^2 S_j^2 / \nu_j^3).$$

Hence show in 36.19 that an approximate F -test of (36.41) may be based on the ratio $(MS)_R / \{(MS)_{RC} + (MS)_{RL} - (MS)_{RCL}\}$, distributed in the variance-ratio form, when (36.41) holds, with $(r-1)$ and ν^* d.fr., where each $c_j^2 = 1$ and $(MS)_j = S_j/\nu_j$. (Satterthwaite (1941) and Box (1954) verify the approximation numerically when the c_j are positive.)

36.8 In 36.18-19, show that in the general r -way balanced cross-classification, only variances with r and $(r-1)$ subscripts can be tested by the UMP similar test of 36.17, but that approximate tests based on Exercise 36.7 can be made for all other variances.

36.9 Verify (36.39) and also (36.44) in Example 36.8.

36.10 In Example 36.3, show that the variance of the unbiased estimator of σ_1^2 is given by

$$\frac{2}{p-1} \left\{ \sigma_1^2 \left(\sigma_1^2 + \frac{2\sigma_\epsilon^2}{n_i} \right) + \frac{(n-1)}{(n-p)n_i^2} \sigma_\epsilon^4 \right\},$$

and hence that the estimator is consistent as $p \rightarrow \infty$, but not if n_i alone $\rightarrow \infty$. (Cf. Exercise 36.1, where test consistency requires that $n_i \rightarrow \infty$.)

(Tukey (1956-7) gives general expressions for variances and covariances of estimators of variances.)

36.11 In Example 36.3, obtain confidence intervals for σ_ϵ^2 and for $\sigma_1^2/\sigma_\epsilon^2$ from the distributions of S_3 and S_2/S_3 respectively. Show that the latter intervals may be partly or wholly below the value zero.

36.12 In 36.25, show that the minimal sufficient statistic for the three parameters has $s = p+1 - \sum_{r=3}^p (r-2)d_r$ components, where d_r is the number of r -tuples of common values among the p values n_i . Show also that $s = 1 + d_1 + 2 \sum_{r=2}^p d_r$.

36.13 Prove (36.50). Show that if and only if the n_i are all equal, its solution with (36.48) gives a unique estimator of σ_1^2 .

36.14 If we denote by z_{g_j} independent χ^2 variables with $2g_j$ d.fr., where the g_j are positive integers, show by resolving the c.f. of $\sum_{j=1}^r c_j z_{g_j}$ into partial fractions that

$$\text{Prob} \left\{ \sum_{j=1}^r c_j z_{g_j} > x \right\} = \sum_{j=1}^r \sum_{s=1}^{g_j} \alpha_{js} \text{Prob} \{z_s > x/c_j\}.$$

Writing $1 - 2itc = y$ in the c.f., show that the constants on the right are given by

$$\alpha_{k, g_k - h} = f_k^{(h)}(0)/h!$$

where

$$f_k(y) = \prod_{\substack{j=1 \\ j \neq k}}^r \{1 + (y-1)\lambda_j/\lambda_k\}^{-g_k}.$$

(Box, 1954)

36.15 If $S_1/(\phi + \lambda)$ and S_2/λ are independent χ^2 variables with f_1, f_2 d.fr. respectively, $M_i = S_i/f_i$, $F = M_1/M_2$ (the variance-ratio statistic), and F_{α, f_2} is the 100α per cent point of the distribution function of F with (f_1, f_2) d.fr., show that if

$$P\{M_2 g(F) \leq \phi\} \doteq \alpha$$

for a monotone increasing function $g(F)$, it should satisfy the conditions

$$g(F_{\alpha, f_2}) = 0,$$

$$g(F) \sim F/F_{\alpha, \infty} \quad \text{as } F \rightarrow \infty.$$

Show that these are satisfied by

$$g_1(F) = (F - F_{\alpha, f_2})/F_{\alpha, \infty}$$

and by

$$g_2(F) = (F/F_{\alpha, \infty}) - 1 + (F_{\alpha, f_2}/F) \{1 - (F_{\alpha, f_2}/F_{\alpha, \infty})\}.$$

(Bulmer (1957) showed that g_1 is a poor, and g_2 a remarkably good, approximation—see also Scheffé (1959).)

CHAPTER 37

THE ASSUMPTIONS OF THE ANALYSIS OF VARIANCE

37.1 When it has been decided that a particular model is appropriate to a given situation, an important problem remains for consideration. Although natural considerations of convenience or technique may dictate that the observations be made on a variable y , it still has to be decided which function of y is to be used for the purpose of the analysis. There is no reason why the quantity measured, rather than some function of it, should be best suited to the assumptions of the model.

There may, indeed, be compelling practical reasons for the consideration of a particular function of y , say $g(y)$ (which may simply be y itself): for example, $g(y)$ may be closely related to the cost or the profitability of a process under investigation. But this implies only that the conclusions of the analysis should finally, for practical purposes, be expressed in terms of $g(y)$; it certainly does not justify the presumption that the model is better satisfied by $g(y)$ than by any other function of y .

37.2 Putting the problem slightly more formally, we may say that a set of observations on y are, equally, a set of "observations" on any well-defined function $g(y)$. The question is, which "observations" $g(y)$ shall we use? Evidently, we must try to determine the function which as nearly as possible satisfies the model. The search for a transformation of this kind was first treated generally by Box and Cox (1964), whose investigation is generally applicable to the linear model (with normal errors) of which the AV Model I in Chapter 35 is a specialization. The reader will observe that the preceding introductory paragraphs are not restricted to the AV context, for the problem is a general one.

Transformations to the normal linear model

37.3 Following Box and Cox (1964), suppose that we observe a dependent variable y and a set of regressor variables x_1, x_2, \dots, x_k ; and that we wish to employ the linear model with normal errors. However, we are not prepared uncritically to assume that we may validly write

$$y = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon};$$

rather, we seek transformations both of y and of each of the x 's so that we have

$$y_{\lambda} = \mathbf{X}_{\mu}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad (37.1)$$

where the components of $\boldsymbol{\epsilon}$ are independently normal with zero means and constant variance σ^2 . In (37.1), $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots)$ indexes the transformation of y within some selected parametric family of transformations, and similarly $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_k)$ indexes the (separate) transformations of the regressors x_1, x_2, \dots, x_k . We are thus generalizing our introductory discussion, where only transformation of y was envisaged.

37.4 By (37.1), the LF is, in logarithmic form,

$$\log L_{\lambda, \mu}(y | \theta, \sigma^2) = -\frac{1}{2}n \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}_\lambda - \mathbf{X}_\mu \theta)'(\mathbf{y}_\lambda - \mathbf{X}_\mu \theta) + \log J_\lambda \quad (37.2)$$

where J_λ is the Jacobian of the inverse transformation from \mathbf{y}_λ (the normally distributed variable in (37.1)) to the actually observed \mathbf{y} . Now, when the LF (37.2) is maximized for given λ, μ , with respect to θ and σ^2 , we find as in 24.28, Vol. 2, that the middle term becomes a constant. If we neglect constants, therefore, we have the conditional maximum for fixed λ, μ ,

$$\log L_{\lambda, \mu}(y | \hat{\theta}, \hat{\sigma}^2) = -\frac{1}{2}n \log \hat{\sigma}_{\lambda, \mu}^2 + \log J_\lambda, \quad (37.3)$$

where $n\hat{\sigma}_{\lambda, \mu}^2 = \mathbf{y}'_\lambda \mathbf{T} \mathbf{y}_\lambda$, say, is the Residual SS, again as in 24.28.

We now need to compute the absolute maximum of the conditional maxima (37.3) over the whole range of λ, μ . This is, even with the aid of electronic computing, a formidable numerical task, except when only one or two transformation indices are involved, e.g. when

- (a) only the dependent variable y is transformed and λ has only one or two components; or
- (b) the same transformation is applied to all of, or a subset of, the regressors, so that μ has only one or two components; or
- (c) λ has a single component as in (a), and (b) holds with only one component in μ .

In cases (b) and (c), numerical plotting of the contours of (37.2) for all λ, μ will generally be necessary. We now confine ourselves to case (a), where only the dependent variable is being transformed. In AV problems, where the regressors are 0-1 variables (cf. 35.9-10) since we are dealing with classified data, this is not a restriction of consequence. In more general regression studies, it implies that we can choose proper forms for the regressor variables before considering transformation of the dependent variable. Box and Tidwell (1962) discuss transformations of the regressors to simpler form (cf. Exercise 37.9); such transformations do not, of course, affect the normality or homoscedasticity of the errors.

37.5 Returning, therefore, to the purpose outlined in our initial discussion, we consider transformations of y alone. In practice, the most useful transformations have been found to be the powers and the logarithm of y , possibly translated by a constant. We therefore consider the family of transformations

$$\begin{aligned} y_\lambda &= (y + \lambda_2)^{\lambda_1}, & \lambda_1 \neq 0, \\ &= \log(y + \lambda_2), & \lambda_1 = 0. \end{aligned} \quad (37.4)$$

To avoid a discontinuity at $\lambda_1 = 0$, we rewrite this equivalently as

$$\begin{aligned} y_\lambda &= \{(y + \lambda_2)^{\lambda_1} - 1\} / \lambda_1, & \lambda_1 \neq 0, \\ &= \log(y + \lambda_2), & \lambda_1 = 0. \end{aligned} \quad (37.5)$$

Tukey (1957b) studied and charted the structural features of the family (37.4) for $\lambda_1 \leq 1$, and Dolby (1963) considered properties of the differential equation which it satisfies, namely

$$\{(y'_\lambda / y''_\lambda)\}' = (\lambda_1 - 1)^{-1}.$$

Healy and Taylor (1962) give tables to facilitate fractional power transformations when $\lambda_2 = 0$ and λ_1 is a multiple of 0.2.

37.6 In (37.3), we now have

$$\log J_\lambda = (\lambda_1 - 1) \sum_{i=1}^n \log(y_i + \lambda_2), \quad (37.6)$$

and (37.3) can be plotted for selected (λ_1, λ_2) for numerical determination of the absolute maximum. An AV must be carried out for each (λ_1, λ_2) used, to obtain the Residual SS in (37.3). In the simplest case when $\lambda_2 = 0$, this can be avoided by equating to zero the first derivative of (37.3) with respect to λ_1 . Using (37.5-6), this gives

$$0 = \frac{\partial \log L_{\lambda_1}(y | \hat{\theta}, \hat{\sigma}^2)}{\partial \lambda_1} = -n \frac{\mathbf{y}'_{\lambda_1} \mathbf{T} \mathbf{u}_{\lambda_1}}{\mathbf{y}'_{\lambda_1} \mathbf{T} \mathbf{y}_{\lambda_1}} + n \lambda_1^{-1} + \sum_{i=1}^n \log y_i, \quad (37.7)$$

where the elements of \mathbf{u} are $\{\lambda_1^{-1} y_i^{\lambda_1} \log y_i\}$.

LR tests of nested hypotheses

37.7 Box and Cox (1964) present some interesting numerical examples of the application of this method of finding a transformation, and of a parallel Bayesian method of analysis which they develop. In addition, they consider the resolution of the maximized LF into three components corresponding to the normality, the homoscedasticity, and the structure of the expectation of y_λ . Their procedure is of general applicability.

Consider sets of constraints C_1, C_2, \dots to be applied successively to a mathematical model, and let $\hat{\lambda}_{(s)}$ be the ML estimator of λ when all of C_1, C_2, \dots, C_s have been applied. $\hat{\lambda}$, without suffix, is the ML estimator when no constraint is imposed. Then, identically for any s ,

$$\begin{aligned} L(y | \hat{\lambda}_{(s)}) &= L(y | \hat{\lambda}) \cdot \frac{L(y | \hat{\lambda}_{(1)})}{L(y | \hat{\lambda})} \cdot \frac{L(y | \hat{\lambda}_{(2)})}{L(y | \hat{\lambda}_{(1)})} \cdot \dots \cdot \frac{L(y | \hat{\lambda}_{(s)})}{L(y | \hat{\lambda}_{(s-1)})} \\ &= L(y | \hat{\lambda}) \cdot l_1 l_2 \dots l_s, \end{aligned} \quad (37.8)$$

where l_p is the LR test statistic for testing the set of constraints $C_1, C_2, \dots, C_{p-1}, C_p$ against the set C_1, C_2, \dots, C_{p-1} (cf. 24.1, Vol. 2). Each of the l_p lies between 0 and 1, and under regularity conditions, $-2 \log l_p$ is asymptotically a non-central χ^2 variable with d.fr. equal to the number of independent constraints upon parameters imposed by C_p (cf. 24.7). When C_p holds, this becomes a central χ^2 variable, and thus $-2 \log l_p$ may be used to test the value of adding C_p to the already imposed C_1, C_2, \dots, C_{p-1} . It should be observed that the l_p are not in general independently distributed, though in particular cases they may be independent under certain hypotheses (cf. Exercises 24.6 and 24.13, and the more general result of Exercise 37.2). The application of the resolution (37.8) to the present problem is left to the reader as Exercise 37.1, since it follows immediately from some results given in Chapter 24.

The purposes of transformation

37.8 The virtue of the ML approach discussed in 37.3-7 is that it requires no prior knowledge of the relationship between y and the regressors, or of the nature of the error distribution of the untransformed y . It starts from the assumption that

there exists some transformation in the family considered for which all the conditions of the linear model, including homoscedasticity and normality of the error distribution, are satisfied. In particular cases, of course, this may not be so; but even then, the ML procedure for choice of the transformation must presumably be an improvement on the uncritical use of y in its original form. It is a striking fact (evidenced by the numerical examples given by Box and Cox (1964)) that this ML transformation is often very close to what is suggested by non-statistical consideration of the nature of the underlying variables. Such consideration should, of course, be undertaken wherever possible as a supplement and guide to the statistical analysis itself.

J. B. Kruskal (1965) gives a computer-based method of finding the monotone transformation of the observations which minimizes the Residual SS (suitably scaled) from an assumed linear model. No parametric family like (37.5) is required; nor is the normality assumption. He uses his method to re-analyse the Box and Cox examples, with several others.

37.9 Other approaches to transforming the data to meet the needs of the linear model have been less ambitious. They seek *either* to normalize the errors *or* to stabilize their variance *or* to remove interactions so that effects are additive; and the hope is general that a transformation which effects one of these aims will at least help towards achieving the others. It is remarkable that this indeed often turns out to be the case, and we shall examine some important instances shortly, but it is over-sanguine to expect this to be always so. It is easy to construct examples where the goals of additivity and homoscedasticity conflict, for if in a two-way cross-classification the expected value of y is additive in row- and column-effects, but the errors are non-normally distributed with variance a function of $E(y)$, any transformation to remove the heteroscedacity will destroy exact additivity, whatever may happen to the non-normality.

We now examine these different types of transformation in turn.

Variance-stabilizing transformations

37.10 Suppose that a statistic t has mean θ and variance, for fixed sample size n ,

$$\text{var } t = D_n^2(\theta). \quad (37.9)$$

To eliminate this dependence of variance on the parameter θ , we seek a function $u(t)$ such that $\text{var } u$ is a constant, c . In general, however, we are unlikely to be able to achieve this precisely, so we ask only that

$$\text{var } \{u(t)\} = c\{1 + O(R^{-1})\} \quad (37.10)$$

where R is some known constant which is large enough for R^{-1} to be negligible. In particular, we may have $R = n$, the sample size. We now assume t to be confined to a neighbourhood of its mean θ . The argument of 10.6-7 then applies, and we have from (10.14) the approximation

$$\text{var } \{u(t)\} = \left\{ \left(\frac{du(t)}{dt} \right)^2 \right\}_{t=\theta} \text{var } t. \quad (37.11)$$

If (37.10) and (37.11) are equated, we have the first-order approximation

$$\left\{ \left(\frac{du(t)}{dt} \right)^2 \right\}_{t=\theta} = \frac{c}{D_n^2(\theta)}. \quad (37.12)$$

Since we are considering only the neighbourhood of θ , we drop the suffix " $t = \theta$," and write θ for t . Thus

$$\frac{du(\theta)}{d\theta} \propto \{D^2(\theta)\}^{-\frac{1}{2}}, \quad (37.13)$$

where we drop the constant c without loss, since this is any case at choice, for multiplication of $u(t)$ by a constant will not affect our purpose of achieving (37.10). We now integrate the equation (37.13), again ignoring the additive constant which results from the indefinite integration without loss, since (37.10) is unaffected. We obtain

$$u(t) \propto \left\{ \int \frac{d\theta}{D_n(\theta)} \right\}_{\theta=t}. \quad (37.14)$$

37.11 Although (37.14) was arrived at through approximation, we can check its validity if the theoretical distribution of t is known by computation of the theoretical variance of $u(t)$ to verify its stability as θ varies—it may be found desirable to modify $u(t)$ to improve stability. Where, on the other hand, we have only observations upon t and no prior knowledge of its distribution or of the parameter θ of that distribution, we cannot even compute $D_n^2(\theta)$ precisely. In such cases, the mean and variance of t in separate groups of observations are calculated, and the latter plotted against the former to give an *estimate* of the relationship (37.9), on which the transformation (37.14) is then based. Here, the approximation is more hazardous, but nevertheless often gives satisfactory results in practice.

Example 37.1

If t has the Poisson distribution, (5.20) shows that mean and variance are equal, so (37.9) is here simply

$$D_n^2(\theta) = \text{var } t = \theta$$

and (37.14) gives

$$u(t) \propto \left\{ \int \theta^{-\frac{1}{2}} d\theta \right\}_{\theta=t} \propto t^{\frac{1}{2}}, \quad (37.15)$$

a simple square-root transformation. To the first order, by (37.11),

$$\text{var } (t^{\frac{1}{2}}) = \left\{ \left(\frac{1}{2} t^{-\frac{1}{2}} \right)^2 \right\}_{t=\theta} \text{var } t = \frac{1}{4}, \quad (37.16)$$

verifying the variance stabilization to this order.

Bartlett (1936) pointed out that variance stabilization could be improved in this case by re-locating t before taking the square root. If we define

$$u_c(t) = (t + c)^{\frac{1}{2}},$$

Bartlett suggested using $c = \frac{1}{2}$. Exercise 37.15 shows that $c = \frac{3}{8}$ is a better choice. The table on the following page gives the variance of $u_c(t)$ as a fraction of its limiting variance as $\theta \rightarrow \infty$, for $c = 0, \frac{1}{2}$ and $\frac{3}{8}$ —the calculations were made by Bartlett (1938) and Anscombe (1948).

THE ADVANCED THEORY OF STATISTICS

θ	Variance of $u_c(t)$ as a fraction of limiting variance		
	$c = 0$	$c = \frac{1}{2}$	$c = \frac{3}{8}$
0	0	0	0
0.5	1.240	0.408	0.717
1.0	1.608	0.640	0.924
2.0	1.560	0.856	0.983
3.0	1.360	0.928	0.999
4.0	1.224	0.960	1.002
6.0	1.104	0.980	
9.0	1.052	0.988	1.001
10.0		0.992	
12.0	1.036	0.992	
15.0	1.024		1.000
20.0			

The inadequacy of the simplest transformation with $c = 0$ is evident for small θ . For θ less than 3, the same comparison is made graphically in Fig. 37.1, adapted from

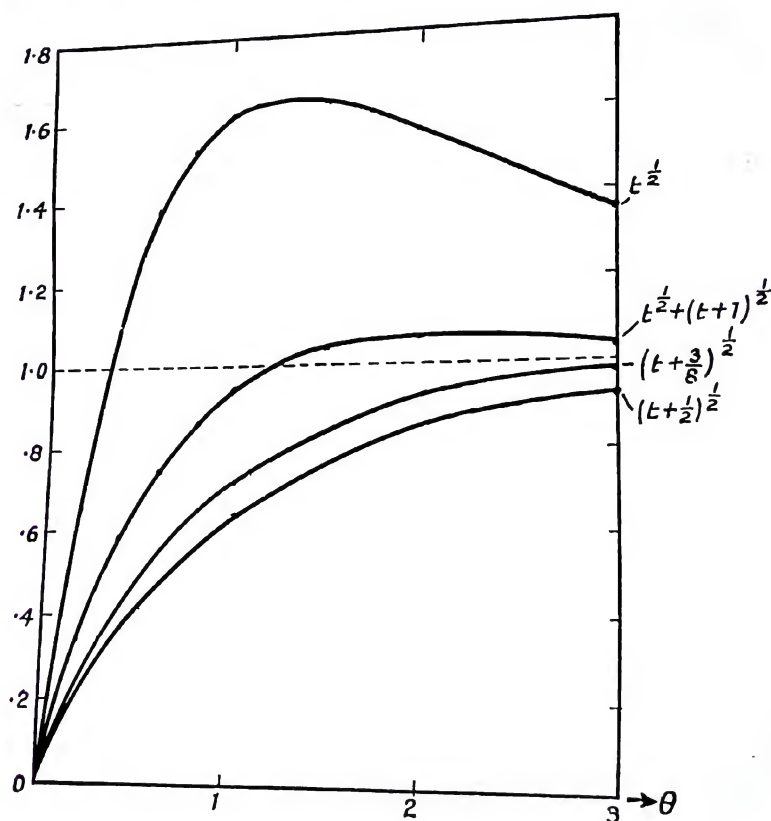


Fig. 37.1

Freeman and Tukey (1950), whose own variance-stabilization proposal,

$$u' = t + (t+1)^{\frac{1}{2}},$$

is more stable than $u_{\frac{1}{2}}(t)$ for $\theta \leq 2$, after which either is adequate. u' is within 6 per cent of stability for $\theta \geq 1$, and seems the best choice (cf. Exercise 37.17).

Example 37.2

In samples from a normal distribution, we know from our earlier work that if t is the sample variance, and σ^2 the population variance, $z = nt/(2\sigma^2)$ is a Gamma variate with parameter $\frac{1}{2}(n-1)$, i.e. the distribution of z is

$$dF(z) = \frac{1}{\Gamma\{\frac{1}{2}(n-1)\}} e^{-z} z^{\frac{1}{2}(n-1)-1} dz$$

(cf., e.g., (11.25) in different notation). The mean and variance of z are therefore each equal to $\frac{1}{2}(n-1)$, and those of t itself are

$$\theta = E(t) = \frac{2\sigma^2}{n} \cdot \frac{1}{2}(n-1) = \sigma^2(n-1)/n,$$

$$D_n^2(\theta) = \text{var } t = \left(\frac{2\sigma^2}{n}\right)^2 \cdot \frac{1}{2}(n-1) = 2\sigma^4(n-1)/n^2,$$

so that here

$$D_n^2(\theta) = 2\theta^2/(n-1).$$

(37.14) gives

$$u(t) \propto \{\int \theta^{-1} d\theta\}_t = \log t, \quad (37.17)$$

and we have arrived at the simple logarithmic transformation. Since

$$\log t = \log z + \log(2\sigma^2/n),$$

the cumulants of $\log t$ and of $\log z$ are identical apart from the constant difference in the mean, and it is easy to see that these cumulants do not depend upon σ^2 at all. The characteristic function of $\log z$ is, writing $p = \frac{1}{2}(n-1)$,

$$\phi(w) = \frac{1}{\Gamma(p)} \int_0^\infty e^{-x} x^{p-1+iw} dx = \Gamma(p+iw)/\Gamma(p).$$

If p is integral (n is odd), this becomes

$$\begin{aligned} \phi(w) &= \frac{(p-1+iw)(p-2+iw) \dots (1+iw) \Gamma(1+iw)}{(p-1)(p-2) \dots 1 \Gamma(1)} \\ &= \Gamma(1+iw) \prod_{s=1}^{p-1} \left(1 + \frac{iw}{s}\right), \end{aligned}$$

so that the cumulant-generating function is

$$\psi(w) = \log \Gamma(1+iw) + \sum_{s=1}^{p-1} \log \left(1 + \frac{iw}{s}\right). \quad (37.18)$$

Now $\Gamma(1-iw)$ is the c.f. of the extreme-value distribution (14.66), with cumulants (cf. Exercise 14.21)

$$\left. \begin{aligned} \kappa_1 &= \gamma \text{ (Euler's constant, } 0.577 \dots), \\ \kappa_r &= (r-1)! \sum_{s=1}^{\infty} s^{-r}, \quad r \geq 2. \end{aligned} \right\} \quad (37.19)$$

Thus the cumulants of $\log z$ obtained from (37.18) are

$$\lambda_r = \left[\frac{\partial^r \psi(w)}{\partial (iw)^r} \right]_{w=0} = (-1)^r \kappa_r + \sum_{s=1}^{p-1} (-1)^{r-1} (r-1)! s^{-r}, \quad (37.20)$$

and substitution of (37.19) into (37.20) gives

$$\left. \begin{aligned} \lambda_1 &= \sum_{s=1}^{p-1} s^{-1} - \gamma, \\ \lambda_r &= (-1)^r (r-1)! \sum_{s=p}^{\infty} s^{-r}, \quad r \geq 2. \end{aligned} \right\} \quad (37.21)$$

Thus, asymptotically, as p increases through the integers,

$$\begin{aligned} \gamma_1 &= \lambda_3 / \lambda_2^{3/2} = -2 \sum_p s^{-3} / \left(\sum_p s^{-2} \right)^{3/2} \sim -\frac{1}{p^2} / \left(\frac{1}{p} \right)^{3/2} = -p^{-1/2}, \\ \gamma_2 &= \lambda_4 / \lambda_2^2 = 6 \sum_p s^{-4} / \left(\sum_p s^{-2} \right)^2 \sim \frac{2}{p^3} / \left(\frac{1}{p} \right)^2 = 2p^{-1}, \end{aligned} \quad (37.22)$$

illustrating the rapidity of the tendency to normality of the distribution of $\log t$.

Bartlett and Kendall (1946) tabulate the mean and variance γ_1 and γ_2 up to $n = 20$ ($p = 9.5$), at which point the asymptotic approximations in (37.22) are adequate.

37.12 The variance-stabilization procedure of **37.10** can be repeated if necessary. Suppose that investigation shows the variance of $u(t)$ to be

$$\text{var } \{u(t)\} = c \left\{ 1 + \frac{p(\theta)}{n} \right\} + o(n^{-1}), \quad (37.23)$$

satisfying (37.10). If we now seek a second transformed variable $v(u)$ such that

$$\text{var } \{v(u)\} = d \{1 + O(n^{-2})\}, \quad (37.24)$$

we have, as at (37.11), the approximation

$$\begin{aligned} \text{var } \{v(u)\} &= \left(\frac{dv(u)}{du} \right)_\theta^2 \text{var } u \\ &= \left(\frac{dv(u)}{du} \right)_\theta^2 c \left\{ 1 + \frac{p(\theta)}{n} \right\} \end{aligned}$$

by (37.23). Using (37.12), this is

$$\begin{aligned} \text{var } \{v(u)\} &= \left(\frac{dv}{du} \right)_\theta^2 \left(\frac{du}{dt} \right)_\theta^2 D_n^2(\theta) \left\{ 1 + \frac{p(\theta)}{n} \right\} \\ &= \left(\frac{dv}{dt} \right)_\theta^2 D_n^2(\theta) \left\{ 1 + \frac{p(\theta)}{n} \right\}. \end{aligned} \quad (37.25)$$

Thus, as at (37.14),

$$v(t) \propto \left\{ \int \left[D_n^2(\theta) \left\{ 1 + \frac{p(\theta)}{n} \right\} \right]^{-1/2} d\theta \right\}_{\theta=t}. \quad (37.26)$$

We have already encountered an instance of this procedure in Hotelling's improved version of Fisher's variance-stabilizing z -transformation at (16.81) (Vol. 1)—cf. Exercises 16.18–19, and Example 37.3 below.

The variance-stabilization procedure could evidently be iterated further if this were necessary.

Exercises 37.4–6 give further applications of a single variance-stabilizing transformation by the method of **37.10**.

Normalizing transformations

37.13 In 6.25-6, we have already examined the Cornish-Fisher method of obtaining a normalizing polynomial transformation; and in 6.27-35, we discussed Johnson's systems of functional transformations to normality.

Curtiss (1943) gives a careful mathematical discussion of the limiting normality of transformations, especially those discussed in our Examples and Exercises. We shall give some examples of the fact mentioned in 37.9, that a transformation designed to achieve one purpose (here, variance stabilization) often also helps to achieve another (here, normalization). In addition, Exercise 37.16 treats the case dealt with in Example 37.1, where the same effect occurs. However, the last of our examples will show that this harmony of purposes is only obtainable by not pressing for optimal achievement in both directions: variance-stabilizing transformations commonly normalize as a by-product, but they do not produce the *optimum* normalization.

Example 37.3

We discussed, in 16.33, Fisher's variance-stabilizing transformation of the correlation coefficient r . The latter was seen at (16.74) to have variance

$$D_n^2(\rho) = \frac{(1-\rho^2)^2}{n} \left(1 + \frac{11\rho^2}{2n} \right) + O(n^{-3}), \quad (37.27)$$

where ρ is the population correlation parameter. (37.14) applied to the leading term of (37.27) gives

$$z(r) = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right)$$

and the variance of z was seen at (16.77) to be

$$\text{var } z = \frac{1}{n-1} \left\{ 1 + \frac{4-\rho^2}{2(n-1)} \right\} + O(n^{-3}), \quad (37.28)$$

depending little upon ρ , so that variance stabilization is good. (16.78) showed that z also has skewness coefficient γ_1 of order $n^{-3/2}$, as against order $n^{-1/2}$ for r ; γ_2 is of order n^{-1} for both.

It seems clear that the variance stabilization symmetrizes, and hence normalizes, as a by-product.

Application of (37.26) here gives

$$z^* = z - (3z+r)/(4n) \quad (37.29)$$

with variance further stabilized at $(n-1)^{-1} + O(n^{-3})$.

Example 37.4

We return to Example 37.2, where we saw at (37.22) that the variance-stabilized logarithmically transformed variable had

$$\gamma_1 = -p^{-1/2}, \quad \gamma_2 = 2p^{-1} \quad (37.30)$$

asymptotically. The untransformed variable is seen from (16.6), with $p = v/2$, to have

$$\gamma_1 = 2p^{-1/2}, \quad \gamma_2 = 6p^{-1}, \quad (37.31)$$

so that the variance stabilization, as by-product, has halved skewness (changing its sign) and reduced kurtosis by a factor of 3.

Example 37.5

In Example 37.2, suppose now that σ^2 is known, and that p is the parameter. We now have $E(z) = \text{var}(z) = p$, and we are back in the same situation as for the Poisson in Example 37.1: (37.14) gives a square-root transformation. The variance stabilization is good, as these values given by Bartlett (1936) show:

p	0	0.5	1.0	2.0	3.0	4.0	9.0	15.0
$\text{var}(p^{\frac{1}{2}})$	0	0.182	0.215	0.233	0.239	0.242	0.247	0.250

This is equivalent to Fisher's approximation to the χ^2 distribution, treated in 16.5-6. It has, by (16.8),

$$\gamma_1 = \frac{1}{2}p^{-1}, \quad \gamma_2 = \frac{3}{16}p^{-2}, \quad (37.32)$$

distinct improvements over (37.31) for the untransformed variable (and better also than (37.30) for its logarithm, which has no virtue in the present case). Again, the variance stabilization has improved the normalization here. But note that the Wilson-Hilferty normalization of 16.7 has, from (16.13), even better γ_1 , of order $p^{-3/2}$, though not such good γ_2 , of order p^{-1} , and (cf. 16.8) gives the better normal approximation. However, it does not stabilize variance at all, as (16.12) shows. Thus the best available normalizing transformation sacrifices variance stability, and the square-root variance stabilizer is a better compromise transformation.

37.14 The reader will see that our discussion of normalization has been couched entirely in terms of skewness and kurtosis. Mathematically, it is taking a good deal for granted to assume that smaller values of γ_1 and γ_2 are equivalent to a closer approach to normality; but we know of no significant example where this assumption misleads us in choosing between normal approximations.

Transformations to additivity

37.15 Although in practice it may be important to search for a scale on which effects are additive (i.e. interactions disappear) or nearly so, relatively little work has been done in this area as compared with normalization and variance stabilization. Some general procedures which have been proposed involve minimization, within a class of transformations, of the value of the test statistic used for the hypothesis that interactions are zero. In the two-way cross-classification, for example, we could minimize S_3 (or S_3/S_R) at (35.65) in the balanced case, or (35.69) in the case of a single observation in each cell. It will be recognized that the test statistic is here being used to carry out a complex estimation procedure, and nothing but intuitive justification has so far been given for this method. Such additivity transformations are sometimes suggested by the analysis of residuals, which we discuss in 37.18-20.

37.16 Other, more specialized, transformations will not be considered in this chapter. We have already (cf. 31.39, Vol. 2) discussed transformation of observations, *via* their ranks, to the standard normal order statistics, or *normal scores*, and in 31.71 mentioned its use in obtaining a distribution-free test for the one-way classification AV situation. The *Probit* and *Logit* transformations of percentages, respectively to normal

and logistic distribution deviates, arise mainly in biological contexts, and are discussed by Finney (1952).

Removal of transformation bias

37.17 Whatever the purpose of a transformation, it often raises problems of presentation when the analysis is complete. In particular, estimators of means or of differences which are unbiased on the transformed scale will not be so if the inverse transformation is made so that results may be presented in "natural" terms (cf., e.g., Exercise 37.15). Adjustments of some kind must be made to remove the bias due to transformation; a general method of bias-reduction was given in 17.10. We now discuss an exact method of removing the bias.

Suppose that u is normally distributed with mean μ and variance σ^2 , and that the functions of u , $(\hat{\mu}, S^2/\nu)$, are jointly sufficient statistics for these parameters, $\hat{\mu}$ being normally distributed with mean μ and variance $\lambda^2 \sigma^2$, and S^2/σ^2 , independent of $\hat{\mu}$, a χ^2 variate with ν d.fr. In practice, we usually have $\lambda^2 = 1/n$ and $\nu = n-1$ where n is sample size. Now consider the function $t(u)$, which in our terms is the *inverse* transformation. Neyman and Scott (1960) (cf. also the succeeding paper by Schmetterer (1960)) showed that if $t(u)$ satisfies the second-order differential equation

$$t''(u) = A + Bt(u)$$

for constants A, B , the unique MV unbiased estimator of the mean of the untransformed variable $\theta = E(t)$ is given by

$$\hat{\theta} = \begin{cases} t(\hat{\mu}) + A(1 - \lambda^2)S^2/(2\nu), & B = 0, \\ \{t(\hat{\mu}) + A/B\} \sum_{r=0}^{\infty} \frac{\Gamma(\frac{1}{2}\nu)}{r! \Gamma(\frac{1}{2}\nu + r)} \left\{ \frac{B(1 - \lambda^2)S^2}{4} \right\}^r - \frac{A}{B}, & B \neq 0. \end{cases}$$

This series converges very rapidly, only a few terms usually being required for adequate accuracy.

It follows that the bias of the crude estimator $t(\hat{\mu})$, which is simply the inverse transformation of $\hat{\mu}$, is

$$E\{t(\hat{\mu}) - \theta\} = \begin{cases} -A(1 - \lambda^2)\sigma^2/2, & B = 0, \\ \{\theta + (A/B)\} [\exp \{-B(1 - \lambda^2)\sigma^2/2\} - 1], & B \neq 0, \end{cases}$$

and its absolute value is always a monotone decreasing function of λ^2 . Since usually $\lambda^2 = 1/n$, the bias will *increase* with sample size.

The following are the most important special cases:

Transformation $u(t)$	Inverse transformation $t(\hat{\mu})$	A	B	Bias $E\{t(\hat{\mu})\} - \theta$	Sign of bias when $\lambda < 1$
$(t+c)^{\frac{1}{2}}$	$\hat{\mu}^2 - c$	2	0	$-(1 - \lambda^2)\sigma^2$	Negative
$\log(t+c)$	$\exp(\hat{\mu}) - c$	c	1	$(\theta + c)[\exp \{- (1 - \lambda^2)\sigma^2/2\} - 1]$	$-\text{sgn}(\theta + c)$
$\arcsin(t)$	$\sin^2(\hat{\mu})$	2	-4	$(\theta - \frac{1}{2})[\exp \{2(1 - \lambda^2)\sigma^2\} - 1]$	$\text{sgn}(\theta - \frac{1}{2})$
$\text{arsinh}(t)$	$\sinh^2(\hat{\mu})$	2	4	$(\theta + \frac{1}{2})[\exp \{-2(1 - \lambda^2)\sigma^2\} - 1]$	$-\text{sgn}(\theta + \frac{1}{2})$

It will be seen that as $\lambda \rightarrow 0$ ($n \rightarrow \infty$), the bias for the square root transformation $\rightarrow -\sigma^2$.

This is the result obtained directly in Exercise 37.15, where $\sigma^2 = \frac{1}{4}$ as at (37.16). The reader may also recall that the bias result for the logarithmic transformation with $c = 0$, $\lambda^2 = 1/n$, was contained in Exercise 18.7, Vol. 2. The other two transformations in the table are those of Exercises 37.4–5.

Analysis of residuals

37.18 A technique which is useful in studying departures from a postulated linear model, and possibly also for suggesting a power transformation to reduce these departures, is given by Anscombe (1961) and by Anscombe and Tukey (1963).

We confine ourselves to the model with a general mean (say, θ_0) which may be written in the form

$$\mathbf{y} = \mathbf{1}\theta_0 + \mathbf{W}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad (37.33)$$

where $\mathbf{1}$ is a vector of units, of order $(N \times 1)$ like $\boldsymbol{\epsilon}$ and \mathbf{y} , \mathbf{W} is a $(N \times p)$ matrix and $\boldsymbol{\theta}$ a $(p \times 1)$ vector. The effect of introducing a general mean θ_0 is (cf. Exercise 19.1) to replace, in the LS estimator of $\boldsymbol{\theta}$, the elements y_i of \mathbf{y} by the deviations $z_i = y_i - \bar{y}$, forming a vector

$$\mathbf{z} = \mathbf{y} - \mathbf{1}\bar{y}, \quad (37.34)$$

and also to replace the elements w_{ij} of \mathbf{W} by the deviations from the column means

$$x_{ij} = w_{ij} - \bar{w}_j, \quad (37.35)$$

forming a matrix \mathbf{X} . Thus we have

$$\mathbf{z}'\mathbf{1} = \mathbf{X}'\mathbf{1} = \mathbf{0}. \quad (37.36)$$

We lose no generality, therefore, by assuming from the beginning that the model is

$$\mathbf{y} = \mathbf{1}\theta_0 + \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

where (37.35) holds. Then the LS estimator of θ_0 is \bar{y} and that of $\boldsymbol{\theta}$ is, by (19.12),

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{z}.$$

We define the matrix $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$, and denote the vector of fitted values by

$$\mathbf{f} = \mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{M}\mathbf{z}$$

and the vector of residuals from the fitted model by

$$\mathbf{r} = \mathbf{z} - \mathbf{f} = (\mathbf{I} - \mathbf{M})\mathbf{z}.$$

By (37.35),

$$\mathbf{r}'\mathbf{1} = \mathbf{0}, \quad (37.37)$$

and since \mathbf{M} is idempotent, we also have

$$\mathbf{M}\mathbf{r} = \mathbf{M}(\mathbf{I} - \mathbf{M})\mathbf{z} = \mathbf{0}. \quad (37.38)$$

37.19 Now suppose that, after fitting the model, we construct a scatter diagram (cf. Example 26.7, Vol. 2) with the fitted values as abscissae and the corresponding residuals as ordinates. By (37.37), the mean ordinate is zero. Further, since \mathbf{M} is symmetric, $\mathbf{f}'\mathbf{r} = (\mathbf{M}\mathbf{z})'\mathbf{r} = \mathbf{z}'\mathbf{M}\mathbf{r} = \mathbf{0}$ by (37.38), so the fitted values are uncorrelated with the residuals, and the regression lines in the scatter diagram are at right angles (cf. 26.9).

Apart from these two general properties of the scatter diagram, we may use its other features to examine how well the assumptions of the fitted model are satisfied.

In particular the homoscedasticity assumption may be checked roughly in terms of the dispersion of the residuals for different sub-ranges of fitted values, and the normality assumption may be checked in the same way, especially so far as skewness is concerned. In each case, an appropriate transformation may be made by the methods discussed earlier in this chapter if the assumption is found to be inadequate.

Perhaps the most interesting use of the scatter diagram, however, is to check additivity assumptions in multi-factor experiments. Non-additivity can manifest itself by evident non-linear (say, quadratic) regression of the residuals upon fitted values.

37.20 The rough visual methods described above can be translated into numerical terms. Anscombe (1961) proposed to use the statistics

$$t_{pq} = \mathbf{r}_p' \mathbf{f}_q$$

where \mathbf{r}_p is the vector of the p th powers of the residuals (e.g. \mathbf{r} defined above is \mathbf{r}_1) and \mathbf{f}_q is the vector of q th powers of the fitted values ($\mathbf{f} = \mathbf{f}_1$). t_{30} and t_{40} are obvious analogues of the usual measures of skewness and kurtosis. t_{21} measures heteroscedasticity, since it essentially gives the covariance of the squared residuals with the fitted values. t_{12} measures non-additivity on the lines indicated at the end of 37.19. In fact, it is very closely related to the statistic S_I used at (35.69) for testing additivity in a two-way cross-classification with one observation per cell—the numerator of S_I is just t_{12}^2 .

37.21 Approximate sampling theory for the t_{pq} (suitably standardized) is developed by Anscombe (1961) and leads to approximations to the power transformations required to remove the corresponding departure from the model's assumptions. In accordance with our discussion of 37.9, there is no guarantee that all these statistics will point to the *same* power transformation, but a general hope that they will not differ by much. In this connexion, it is interesting that Box and Cox (1964) (see also the discussion of their paper by Anscombe) expressed the ML solution (37.7) for the power transformation approximately in terms of the t_{pq} with $p+q = 3$ and 4. In essence, the ML estimator carries out a kind of averaging process between the various power transformations suggested by the individual measurements of heteroscedasticity, non-normality and non-additivity. It is not the least of its virtues that the ML approach automatically effects what might otherwise be a difficult compromise to make.

The robustness of AV procedures

37.22 We first consider estimation of the parameters in AV problems. Where Model I AV is concerned, LS estimation theory and its optimum properties (set out in 19.4–9, Vol. 2) does not at all involve the assumption of normality for the errors. Thus all estimates remain valid, and so do their estimated variances, in face of non-normality: LS estimation is distribution-free to this extent. The normality assumption was required in 24.27–37 for hypothesis-testing and interval estimation purposes only.

Further, even if the basic LS model (19.8) is incorrect in respect of its assumption of uncorrelated, homoscedastic errors, this will not bias the LS estimator (19.12), for (19.13–14) hold so long as the errors have zero means. But the MV properties are lost

in this case, since they now pass to the true LS estimator (19.59). Thus heteroscedasticity and correlation of the errors merely reduce efficiency without importing bias. For model II AV, and the other models which are considered in Chapter 36, it is easy to see that the expectations of Mean Squares are unaffected by the failure of the normality assumption, so that estimators of variance components remain unbiased in the presence of non-normality of the various random variables. However, the variances of these estimators (e.g. in the simplest case given in Exercise 36.10) are radically changed by non-normality, because they are no longer simplified by the special relations between the cumulants of the normal distribution.

37.23 So far as tests (and the corresponding interval estimators) are concerned, we noticed in 31.2-9, Vol. 2, the outstanding general feature of the effects of non-normality upon normal-theory procedures: tests on means are robust, while tests on variances are not. This generalization leads us to expect that tests and interval estimates in Model I, which is essentially concerned with means, will be relatively robust to non-normality; and that those in Model II and other AV models, which are concerned with variances, will not be robust. We treat robustness to non-normality in detail in 37.24-35, but here remark that these statements have been substantially justified in some detailed investigations, very fully summarized in the final chapter of Scheffé (1959); an earlier review by Cochran (1947) may also be consulted.

These investigations (e.g. Horsnell (1953); Box (1954)) also showed that in Model I the effects of heteroscedasticity of errors can be large in general, but are not serious when equal frequencies are used in all cells of the classification. (We have previously countered this effect in simpler form in 21.24 (cf. also 31.4).) The practical implication is that on grounds of robustness alone equal cell-frequencies should be used wherever possible when the observations are designed. As a happy side-effect, computations are made much easier by this conclusion. Further, this robustness to heteroscedasticity in the balanced case permits us to make a simple approximate AV of cell means when all frequencies are unequal (cf. Exercises 37.7-8).

The effects of stochastic dependence among the errors can be extreme (Box, 1954). This recalls a general point made in 36.39-43, that randomization methods of allocating experimental units (which may obviously introduce some dependencies among the errors) should be taken into account in the analysis. We shall return to these methods in Chapter 38.

Robustness to non-normality in the linear model

37.24 An interesting approach due to Box and Watson (1962), following earlier work by Box and Andersen (1955), throws some general light upon the reasons for varying degrees of robustness to non-normality.

We return to the linear model with general mean, defined in 37.18. The SS attributable to the fitted model there is

$$S_0 = \mathbf{f}'\mathbf{f} = \mathbf{z}'\mathbf{M}\mathbf{z},$$

and the Residual SS is

$$S_R = \mathbf{r}'\mathbf{r} = \mathbf{z}'(\mathbf{I} - \mathbf{M})\mathbf{z}.$$

When the errors are normally distributed, the LR test of $\theta = 0$ is based on

$$F = (S_0/p) \{S_E/(N-p-1)\},$$

distributed in the variance-ratio form with d.fr. $\nu_1 = p, \nu_2 = N-p-1$. The test can equivalently be carried out on

$$W = \frac{S_0}{S_0 + S_R} = \frac{\mathbf{z}'\mathbf{M}\mathbf{z}}{\mathbf{z}'\mathbf{z}}, \quad (37.39)$$

for since

$$W = \left(1 + \frac{\nu_2}{\nu_1 F}\right)^{-1}, \quad (37.40)$$

it is a monotone increasing function of F . In the normal case, when $\theta = 0$, $1/F$ has the variance-ratio distribution with $(N-p-1, p)$ d.fr. and W is the Beta variable, with parameters $\{\frac{1}{2}(N-p-1), \frac{1}{2}p\}$, obtained from $1/F$ by the transformation in 16.19.

We now study the distribution of W in the general (non-normal) case. Its denominator is invariant under permutation of the elements of \mathbf{z} . We therefore first consider this *permutation distribution* (cf. 31.16) of W . If the joint distribution of the elements of \mathbf{z} is symmetric in its N arguments, as will be so in particular when the errors are independently and identically distributed, each permutation of them has probability $(N!)^{-1}$.

Once we have obtained the mean and variance of W in this permutation distribution, say $E_P(W)$ and $V_P(W)$, we shall be able to obtain the unconditional mean $E(W)$ and variance $V(W)$ from them if we know the parent distribution from which \mathbf{z} was sampled.

37.25 Since $\mathbf{z}'\mathbf{M}\mathbf{z}$ is a scalar,

$$\mathbf{z}'\mathbf{M}\mathbf{z} = \text{tr}(\mathbf{z}'\mathbf{M}\mathbf{z}) = \text{tr}(\mathbf{M}\mathbf{z}\mathbf{z}')$$

where we commute under the trace operator. Thus (37.39) gives

$$\mathbf{z}'\mathbf{z}E_P(W) = E_P\{\text{tr}(\mathbf{M}\mathbf{z}\mathbf{z}')\} = \text{tr}\{\mathbf{M}E_P(\mathbf{z}\mathbf{z}')\}.$$

Now

$$E_P(\mathbf{z}\mathbf{z}') = \frac{1}{N(N-1)}\mathbf{z}'\mathbf{z}(\mathbf{N}\mathbf{I} - \mathbf{1}\mathbf{1}'), \quad (37.41)$$

since

$$E_P(z_j^2) = \mathbf{z}'\mathbf{z}/N, \quad E_P(z_j z_l) = -\mathbf{z}'\mathbf{z}/\{N(N-1)\}$$

for $j \neq l$. Substitution of (37.41) gives

$$\begin{aligned} E_P(W) &= \frac{1}{N(N-1)} \text{tr}\{\mathbf{M}(\mathbf{N}\mathbf{I} - \mathbf{1}\mathbf{1}')\} \\ &= \frac{1}{N-1} \text{tr}(\mathbf{M}) \\ &= p/(N-1), \end{aligned} \quad (37.42)$$

since $\mathbf{M}\mathbf{1}\mathbf{1}' = 0$ by (37.35) and $\text{tr}(\mathbf{M}) = p$ from 19.9.

(37.42) shows that $E_P(W)$ does not depend upon \mathbf{X} or upon \mathbf{z} . Thus, whatever the distribution of the errors, say f , the unconditional mean of W will also be

$$E(W) = E_f\{E_P(W)\} = p/(N-1). \quad (37.43)$$

In particular, this will hold in the normal case, so that the mean of W is completely robust to departures from normality.

37.26 To obtain $V_P(W)$, we first find $E_P(W^2)$. Writing M_{rs} for the elements of \mathbf{M} , (37.44)

where we now always sum over all possible unequal values of the subscripts. Squaring (37.44) and taking expectations over all permutations of the elements of \mathbf{z} , we find

$$E_P\{(\mathbf{z}'\mathbf{M}\mathbf{z})^2\} = E_P(z_r^4) \sum M_{rr}^2 + E_P(z_r^3 z_s) 4 \sum M_{rr} M_{rs} \\ + E_P(z_r^2 z_s^2) (2 \sum M_{rs}^2 + \sum M_{rr} M_{ss}) \\ + E_P(z_r z_s^2 z_t) (4 \sum M_{rs} M_{st} + 2 \sum M_{rr} M_{st}) \\ + E_P(z_r z_s z_t z_u) \sum M_{rs} M_{tu}. \quad (37.45)$$

We now write s_r for the r th power-sum of the z 's, as at (12.8) (so that $s_1 = 0$ by (37.35), while $\mathbf{z}'\mathbf{z} = s_2$), and use the relations between the augmented symmetric functions and power-sums to evaluate the expectations in (37.45), using (12.9). From the weight 4 section of Appendix Table 10, we find

$$\left. \begin{aligned} NE_P(z_r^4) &= s_4, \\ N^{(2)} E_P(z_r^3 z_s) &= -s_4, \\ N^{(2)} E_P(z_r^2 z_s^2) &= s_2^2 - s_4, \\ N^{(3)} E_P(z_r z_s^2 z_t) &= 2s_4 - s_2^2, \\ N^{(4)} E_P(z_r z_s z_t z_u) &= 3s_2^2 - 6s_4. \end{aligned} \right\} \quad (37.46)$$

Furthermore, we may express all the sums involving the M_{rs} in (37.45) in terms of $m = \sum_{r=1}^N M_{rr}^2$, using the idempotency of \mathbf{M} , the value p of its trace, and the fact that $\mathbf{M}\mathbf{1} = \mathbf{0}$ by (37.35). These relations are:

$$\left. \begin{aligned} \sum M_{rr} M_{rs} &= -m, \\ \sum M_{rs}^2 &= p - m, \\ \sum M_{rr} M_{ss} &= p^2 - m, \\ \sum M_{rs} M_{st} &= 2m - p, \\ \sum M_{rr} M_{st} &= 2m - p^2, \\ \sum M_{rs} M_{tu} &= p^2 + 2p - 6m. \end{aligned} \right\} \quad (37.47)$$

Substituting (37.46-7) into (37.45), and writing

$$k_2 = s_2/(N-1), \quad k_4 = \{N(N+1)s_4 - 3(N-1)s_2^2\}/(N-1)^{(3)},$$

the k -statistics of the y 's by (12.28), we find, on dividing by $(\mathbf{z}'\mathbf{z})^2 = s_2^2$, subtracting $\{E_P(W)\}^2$ and simplifying,

$$V_P(W) = \frac{2p(N-p-1)}{(N+1)(N-1)^2} + \frac{k_4}{k_2^2(N-1)^2} \left\{ m - \frac{p^2}{N} - \frac{2p(N-p-1)}{N(N+1)} \right\}. \quad (37.48)$$

37.27 Unlike (37.42), (37.48) depends on \mathbf{X} through m , and upon \mathbf{z} through the ratio k_4/k_2^2 . Because we found $E_P(W) = E(W)$, we see that

$$V(W) = E(W^2) - \{E_P(W)\}^2,$$

$$V_P(W) = E_P(W^2) - \{E_P(W)\}^2,$$

while
so that if f is the distribution of the errors,

$$V(W) = E_f\{V_P(W)\}. \quad (37.49)$$

The unconditional variance of W thus depends essentially on $E(k_4/k_2^2)$ in (37.48). In the normal case,

$$E(k_4/k_2^2) = E(k_4)/E(k_2^2) = 0,$$

since k_2 is distributed independently of k_4/k_2^2 . (k_2 is a complete sufficient statistic for, and k_4/k_2^2 is distributed free of, the scale parameter, and Exercise 23.7 applies.) Thus the first term on the right of (37.48) is the normal-theory unconditional variance, and we rewrite the result as

$$V_P(W) = \{V(W)\}_{\text{Normal}} \left[1 + C_y \frac{(N+1)}{2p(N-p-1)} \left\{ m - \frac{p^2}{N} - \frac{2p(N-p-1)}{N(N+1)} \right\} \right], \quad (37.50)$$

where $C_y = k_4/k_2^2$.

37.28 Exercise 37.10 asks the reader to show that m can be expressed in terms of the k -statistic ratios $(k_4/k_2^2)_i$ and $\{k_{22}/(k_{20}k_{02})\}_{ij}$ of the p regressors x . Using that result, (37.50) may be written in the form

$$V_P(W) = \{V(W)\}_{\text{Normal}} \left[1 + \frac{(N-3)}{2N(N-1)} C_y C_x \right], \quad (37.51)$$

where

$$\begin{aligned} C_x &= \frac{N(N^2-1)}{p(N-p-1)(N-3)} \left\{ m - \frac{p^2}{N} - \frac{2p(N-p-1)}{N(N+1)} \right\} \\ &= \frac{(N-2)}{p(N-p-1)} \left\{ \sum_{i=1}^p \left(\frac{k_4}{k_2^2} \right)_i + \sum_{i \neq j=1}^p \left(\frac{k_{22}}{k_{20}k_{02}} \right)_{ij} \right\}, \end{aligned} \quad (37.52)$$

and

$$\{V(W)\}_{\text{Normal}} = \frac{2p(N-p-1)}{(N+1)(N-1)^2}. \quad (37.53)$$

(37.52) shows that C_x is a multivariate generalization of the univariate kurtosis ratio k_4/k_2^2 , to which it reduces when $p = 1$. It has zero mean in the normal case, by the argument given for k_4/k_2^2 in 37.27.

37.29 The permutation distribution's moments (37.42) and (37.51) permit us to fit a continuous distribution to the discrete permutation distribution of W in the manner of 31.47, by choosing a Beta distribution with the same mean and variance. Since both W and the fitted distribution are on the range $(0, 1)$, and we know (cf. 37.24) that this distribution holds exactly for the unconditional distribution of W in normal samples, there is a reasonable hope of obtaining a good approximation to the general permutation distribution.

The mean and variance of a Beta distribution with parameters $\frac{1}{2}\nu_2, \frac{1}{2}\nu_1$ are (from Example 2.8)

$$\begin{aligned} \mu'_1 &= \frac{\nu_1}{\nu_1 + \nu_2}, \\ \mu_2 &= \frac{2\nu_1\nu_2}{(\nu_1 + \nu_2 + 2)(\nu_1 + \nu_2)^2}, \end{aligned}$$

whence

$$\left. \begin{aligned} v_2 &= v_1(1 - \mu'_1)/\mu'_1, \\ v_1 &= 2\mu'_1(\mu'_1 - \mu'^2_1 - \mu_2)/\mu_2. \end{aligned} \right\} \quad (37.54)$$

μ'_1 and μ_2 are to be equated to $E_P(W)$ and $V_P(W)$ respectively.

Since v_2/v_1 is a function of μ'_1 only by (37.54), and $E_P(W)$ is constant at $p/(N-1)$, we see that v_2 will require adjustment by the same factor as v_1 . Substituting $E_P(W)$ for μ'_1 and $V_P(W)$ for μ_2 in (37.54), we find

$$v_1 = p \left\{ 1 + \frac{(N+1)c}{N-1-2c} \right\}^{-1} \quad (37.55)$$

where c is the "correction factor" in (37.51), i.e.

$$c = \frac{N-3}{2N(N-1)} C_y C_x. \quad (37.56)$$

It follows that

$$v_2 = (N-p-1) \left\{ 1 + \frac{(N+1)c}{N-1-2c} \right\}^{-1} \quad (37.57)$$

also. If either C_y or $C_x = 0$, $c = 0$ and normal theory holds for $V_P(W)$, to our approximation, whatever the underlying distribution of the errors.

37.30 (37.49) and (37.51) show that the unconditional variance $V(W)$ is simply $V_P(W)$ with $C_y C_x$ replaced by $\frac{E(C_y C_x)}{f}$. With this modification, the approximation of 37.29 holds for the unconditional distribution of W as well as for its permutation distribution.

37.31 Since the approximating d.f.r. defined at (37.55-7) depend essentially on the correction factor c , it is of interest to find bounds for its constituents. Exercise 37.11 is to show that

$$p^2/N \leq m \leq p(N-1)/N, \quad (37.58)$$

and hence, from (37.52),

$$-2 \leq C_x(N-3)/(N-1) \leq N-1. \quad (37.59)$$

We see that, if these bounds for C_x are attained or nearly so, the correction factor c at (37.56) will be of order N^{-1} near the lower bound and of order N^0 near the upper bound. This will determine, at least in large samples, whether the correction factor is negligible or not, i.e. whether the distribution of W is robust or not. Since we have seen at (37.52) that the deviation of C_x from zero is a measure of multivariate normality for the x 's, we may say that normality of the regressors produces robustness to non-normality in W .

Robustness to non-normality in one-way AV

37.32 We now apply the general results for the linear model in 37.22-31 to the particular case of a one-way classification. Since there are $(p+1)$ parameters (θ_0, θ) in (37.36), we suppose that there are $(p+1)$ groups in the classification—the re-parameterized form of the model in the later part of Example 35.1 will then be non-singular, since we do not have a surplus parameter producing singularity.

Suppose that there are n_j observations in the j th group, $\sum_{j=1}^{p+1} n_j = N$. Each of the p regressors must now simply indicate whether an observation does or does not fall into a particular group; membership of the $(p+1)$ th group is implied by non-membership of all the others. Ordinarily, we should define the regressors as 0-1 variables for this purpose, but we must satisfy (37.36) and therefore instead define

$$x_{ij} = \begin{cases} 1 - \frac{n_j}{N} & \text{when the observed } y \text{ falls into the } j\text{th group,} \\ -\frac{n_j}{N} & \text{when it does not,} \end{cases} \quad (37.60)$$

for $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, p$. Then $\sum_{i=1}^N x_{ij} = 0$ as required for each j .

In this case we know from Example 35.1 that the SS attributable to the fitted model is, temporarily using bars instead of dot suffixes for averages,

$$\mathbf{z}'\mathbf{M}\mathbf{z} = \mathbf{y}'\mathbf{M}\mathbf{y} = \sum_{j=1}^{p+1} n_j (\bar{y}_j - \bar{y})^2 \equiv \sum_j \frac{\left(\sum_{i=1}^{n_j} y_{ij}\right)^2}{n_j} - \frac{\left(\sum_i \sum_j y_{ij}\right)^2}{N}, \quad (37.61)$$

and from (37.61) we see by considering the coefficient of y_{ij}^2 that $M_{rr} = \frac{1}{n_j} - \frac{1}{N}$ for every member of the j th group. Thus

$$m = \sum_{r=1}^N M_{rr}^2 = \sum_{j=1}^{p+1} n_j \left(\frac{1}{n_j} - \frac{1}{N}\right)^2 = \sum_{j=1}^{p+1} \frac{1}{n_j} - \frac{2p+1}{N},$$

and (37.52) becomes

$$\frac{N-3}{N-1} C_X = \frac{N(N+1)}{p(N-p-1)} \left\{ \sum_{j=1}^{p+1} \frac{1}{n_j} - \frac{(p+1)^2}{N} \right\} - 2, \quad (37.62)$$

which can be substituted into (37.51) to get $V_P(W)$, first given by Welch (1937). Evidently, the value of C_X will depend critically on how the N observations are allocated to the $(p+1)$ groups.

37.33 If $n_j = N/(p+1)$, so that all groups have the same frequency, (37.62) becomes

$$\frac{N-3}{N-1} C_X = -2, \quad (37.63)$$

attaining the lower bound in (37.59). (37.56) is therefore

$$c = -C_y/N,$$

and the multiplier to be applied to p and $(N-p-1)$ in (37.55) and (37.57) is

$$\left(1 + \frac{2C_y}{N(N-1)}\right) / \left(1 - \frac{C_y}{N}\right) \sim 1 + \frac{C_y}{N},$$

negligible in large samples.

As is indicated by the fact that zero lies between the bounds in (37.59), this negligible correction can actually be reduced to zero (cf. Exercise 37.12) by making group frequencies unequal in a certain way, but in general it seems unwise to produce this slight

further increase in robustness to non-normality at the expense of losing the robustness to heteroscedasticity in the balanced case (cf. 37.21).

37.34 As an extreme contrast to the equal-frequencies case treated in 37.33, consider the case $n_1 = n_2 = \dots = n_p = 1$, $n_{p+1} = N-p$. (37.62) now becomes

$$\frac{N-3}{N-1} C_x = N-1 - \frac{N+1}{N-p}, \quad (37.64)$$

which is very near the upper bound in (37.59) if p/N is small. (37.56) is

$$c = \frac{C_y}{2N} \left(N-1 - \frac{N+1}{N-p} \right) \sim \frac{C_y}{2} \left(1 - \frac{1}{N} \right),$$

and the multiplier in (37.55) and (37.57) is approximately $1 + \frac{1}{2} C_y$, indicating extreme non-robustness.

The point of most interest about this opposite extreme emerges only when we calculate the SS (37.61), which is

$$\mathbf{y}' \mathbf{M} \mathbf{y} = \sum_{j=1}^p (y_{1j} - \bar{y})^2 + (N-p)(\bar{y}_{p+1} - \bar{y})^2. \quad (37.65)$$

Now it is clear that as N increases, \bar{y}_{p+1} and $\bar{y} = \left\{ \sum_{j=1}^p y_{1j} + (N-p)\bar{y}_{p+1} \right\} / N$ will differ negligibly. Thus, to the first order in N , (37.65) is simply

$$\sum_{j=1}^p (\bar{y}_{1j} - \bar{y})^2 \sim \sum_j (y_{1j} - y.)^2 + p(y. - \bar{y}_{p+1})^2,$$

where $y. = \sum_{j=1}^p y_{1j} / p$.

The F -test in 37.22 will be based on the ratio of this to

$$\sum_{i=1}^p \sum_{j=1}^{p+1} (y_{ij} - \bar{y})^2 - \sum_{j=1}^p (y_{1j} - \bar{y})^2 = \sum_{i=1}^{N-p} (y_{i,p+1} - \bar{y})^2 \sim \sum_{i=1}^{N-p} (y_{i,p+1} - \bar{y}_{p+1})^2.$$

Thus

$$F \sim \frac{\left\{ \sum_{j=1}^p (y_{1j} - y.)^2 + p(y. - \bar{y}_{p+1})^2 \right\} / p}{\sum_{i=1}^{N-p} (y_{i,p+1} - \bar{y}_{p+1})^2 / (N-p-1)}. \quad (37.66)$$

Apart from the term which compares $y.$ and \bar{y}_{p+1} in the numerator, and the corresponding extra degree of freedom there, (37.66) is the F -statistic for testing the equality of variances in two normal populations, from samples of size p and $N-p$ (cf. Exercise 23.14). In the light of the results of this section, it is easy to understand the extreme non-robustness of the latter test, referred to in general terms in 37.21 and in more detail in 31.6-8, where essentially the same correcting multiplier $(1 + \frac{1}{2} C_y)$ was justified directly for tests on variances generally.

Robustness to normality in balanced classifications

37.35 In the balanced one-way classification of 37.33, (37.63) and (37.52) show that

$$m = \sum_{r=1}^N M_{rr}^2 = p^2 / N. \quad (37.67)$$

The lower bound of (37.59) will be attained, and a negligible correction for normality will result as in 37.33, whenever (37.67) is satisfied, and in particular whenever

$$M_{rr} = p/N \quad (37.68)$$

for all r , i.e. whenever the diagonal elements of $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ are all equal. When a linear model satisfies (37.68) it is said to be *quadratically balanced*. It is easy to see from considerations of symmetry alone that (37.68) holds for any cross-classification with equal frequencies in every cell (previously called "balanced") and for hierarchical classifications with equal frequencies at each stage of the hierarchy. Atiqullah (1962), using a different method, derived this as an asymptotic result for the unconditional distribution of W ; his results also apply to other F -tests than the overall test of $\theta = 0$, which is the only one considered here.

In analysis of covariance in the one-way classification, with equal frequencies, Atiqullah (1964) showed that in accordance with the conclusion of 37.31, the extent of non-normality of the concomitant variable determines the robustness of the F -test to non-normality of the errors; robustness to other departures from assumptions is also considered.

Distribution-free methods in AV

37.36 The study of robustness of the standard AV methods leads us naturally to enquire whether other, completely robust, methods of analysis can be found. In other words, are there distribution-free methods for AV problems?

We have already seen, in 31.70-4, that so far as the one-way classification is concerned, the answer is in the affirmative: distribution-free tests exist for the equality of the location parameters of k samples from any continuous populations otherwise of the same form. The test may be based on the ranks themselves, using the test statistic (31.150), or on the normal scores $E(S, n)$ (cf. 31.71, Vol. 2). These tests are completely robust for any continuous distribution and have very high asymptotic relative efficiencies against normal location-shift alternatives (cf. 31.71). Another test, against ordered alternatives, was given in 31.72-4 (cf. 35.66 and Exercise 35.15).

37.37 Further, the permutation distribution of W in the general linear model, discussed in 37.24-31, is distribution-free in the same way as permutation tests were in Chapter 31; the test of $\theta = 0$ holds as an approximate test for the symmetry of \mathbf{z} in its arguments whatever the underlying distribution of the errors. However, this test of $\theta = 0$ does not carry us very far into AV except for the one-way classification, as we saw in 37.32-5. We now have to consider how far distribution-free methods may be of use in more complex AV situations, where we wish to test main effects, interactions, etc.

Two-way cross-classification: permutation test

37.38 Consider the simplest two-way cross-classification, with one observation per cell. Suppose that there are r rows and c columns, so that there are $n = rc$ observations in all, and that we wish to test column- (or row-) effects. In the spirit of our discussions of 31.21 and 31.39, the most natural procedure would be to replace the n observations y_{ij} by their ranks, or by some other set of conventional numbers such

as the normal scores, and carry out the usual AV tests upon these. It is difficult to make any progress with the distribution theory of this procedure, since there is no set of equiprobable permutations from which to deduce results.

37.39 However, we may develop a permutation test for column-effects from which, as we shall see, distribution-free tests will emerge.

The usual AV SS for testing column-effects, say $S_C = r \sum_{j=1}^c (y_{.j} - y_{..})^2$, is invariant under the addition of a constant to each observation in any row. If we take the mean of each row to be zero, we make the rows SS, say $S_R = c \sum_{i=1}^r (y_{i.} - y_{..})^2$, equal to zero since $y_{i.} = y_{..} = 0$, and also reduce S_C to

$$S_C = r \sum_j y_{.j}^2.$$

(37.69)

Now if $S = \sum_i \sum_j (y_{ij} - y_{..})^2$, the standard F -statistic is

$$F = \frac{S_C / (c-1)}{(S - S_R - S_C) / \{(r-1)(c-1)\}},$$

with d.f. $\nu_1 = c-1$ and $\nu_2 = (r-1)(c-1)$. Like S_C , $S - S_R$ is invariant under arbitrary addition of constants to the rows, and therefore F is. We may therefore without loss of generality put $S_R \equiv 0$ and $F = \frac{(r-1)S_C}{S - S_C}$. Its Beta transform (cf. (37.40)) is

$$\begin{aligned} W &= \left(1 + \frac{\nu_2}{\nu_1 F}\right)^{-1} = \frac{S_C}{S} \\ &= \frac{\sum_{j=1}^c \left(\sum_{i=1}^r y_{ij}\right)^2}{r \sum_i \sum_j y_{ij}^2} = \frac{1}{r} \left\{ 1 + \frac{2U}{(c-1) \sum_{i=1}^r (k_2)_i} \right\}, \end{aligned} \quad (37.70)$$

where

$$U = \sum_{i=1}^r \sum_{l=1, l \neq i}^r \sum_{j=1}^c y_{ij} y_{lj}$$

and $(k_p)_i$ is the p th k -statistic of the y -values in the i th row, $y_{i1}, y_{i2}, \dots, y_{ic}$.

37.40 We can now find the moments of (37.70), just as we did those of (37.40), under a hypothesis of symmetry. Here, the hypothesis is that there are no column-effects in the classification. Because of the invariance of F , and hence W , to constant additions to the rows, we need make no assumption at all about row-effects. The hypothesis implies that in each of the r rows of the classification separately, every one of the $c!$ permutations of the y_{ij} ($j = 1, 2, \dots, c$) is equiprobable. In all, therefore, there are $(c!)^r$ equiprobable arrangements of the y_{ij} under the hypothesis. The case $c = 2$ corresponds to Fisher's test of bivariate symmetry discussed in 31.78. Pitman (1938), whose paper should be consulted for details, found the first four

moments of U , and thence of W . He found

$$E_P(W) = \frac{1}{c}. \quad (37.71)$$

At (37.71), just as at (37.42-3), the mean of the permutation distribution is the same whatever the observations, and therefore coincides with the normal-theory result. The variance, as at (37.52), is more complicated, being (cf. Exercise 37.13)

$$V_P(W) = \frac{2}{r^2(c-1)} \left[1 - \frac{\sum_i (k_2^2)_i}{\{\sum_i (k_2)_i\}^2} \right], \quad (37.72)$$

with more complicated expressions for the higher moments.

The normal theory variance is, as in 37.29,

$$\{V(W)\}_{\text{Normal}} = \frac{2\nu_1\nu_2}{(\nu_1 + \nu_2 + 2)(\nu_1 + \nu_2)^2} = \frac{2(r-1)}{r^2\{r(c-1) + 2\}} \quad (37.73)$$

and, as at (37.49), this is the expectation of (37.72) under normality. Solving (37.72-3), we find that

$$E \left[\frac{\sum_i (k_2^2)_i}{\{\sum_i (k_2)_i\}^2} \right] = \frac{c+1}{r(c-1)+2} = \frac{1 + \frac{2}{c-1}}{r + \frac{2}{c-1}} \quad (37.74)$$

in the normal case.

The d.f. of the F -test can be adjusted as in 37.29—Exercise 37.14 gives an instance. Pitman (1938) showed that when the mean and variance are made to agree with those of the Beta distribution in this way, the third and fourth moments also generally show good agreement.

37.41 The most interesting special cases of W for our present purposes are those in which the observations y_{ij} are replaced by conventional numbers, so that the test is made completely distribution-free. Instead of the procedure outlined in 37.38, we shall now replace the y_{ij} in each row separately by a set of conventional numbers, e.g. their ranks or the corresponding normal scores. If we use the same set in each row, an immediate consequence is that the $(k_p)_i$ are identical for all values of i . Thus

$$\frac{\sum_i (k_2^2)_i}{\{\sum_i (k_2)_i\}^2} = \frac{rk_2^2}{(rk_2)^2} = \frac{1}{r} \quad (37.75)$$

simply. If (37.75) is compared with (37.74), it is seen that they differ negligibly if either r or c is not too small. The distribution-free test statistic using the same set of conventional numbers to replace the observations in each row will then have approximately the same distribution as the normal theory test. In particular, this is true if the ranks or the normal scores are used in place of the observations.

It should be noted that when conventional numbers are used in the test, there is no need to put $y_{i.} = 0$, for S_R will be $\equiv 0$ in any case. Of course $y_{..}$ (a constant) must be restored to S_C and S in this case, as in Exercise 37.14.

More complex classifications

37.42 Thus there are distribution-free AV tests of column- (or, by transposition, of row-) effects irrespective of the existence of row- (column-) effects, in a two-way cross-classification with one observation per cell. The simplicity of this situation arises because, as we saw in 37.39, the total SS (neglecting the general mean) has only three components (S_R , S_C , $S - S_R - S_C$) of which one (S_R) may be rendered identically zero by suitable choice of origin. F is then a ratio of the only two random variables in the problem. When conventional numbers are used, S too becomes a constant, so that the Beta transform W is just a constant multiple of the only remaining random variable, S_C .

As soon as we begin to consider generalizing the permutation test, this simplicity disappears. Even in the balanced two-way cross-classification with more than one observation per cell, the total SS has a further (Interactions) component; for a three-way cross-classification with one observation per cell, also, an extra component appears. In consequence, the permutation distribution of the F - or W -statistic for column-effects in each of these cases (and *a fortiori* in more complex situations) will be difficult, and this is presumably why these tests have not been developed.

37.43 An alternative method of generalization would be to consider the analogue of one of the distribution-free statistics in the more complex situations. For example, in Exercise 37.14, where ranks are used, W is a multiple of T , essentially the variance of the column total ranks. The distribution of this variance might be obtained for the balanced two-way cross-classification, and possibly also for the unequal-frequencies case; and for the three-way ($r \times c \times l$) cross-classification under the $(c!)^{rl}$ equiprobable permutations of the column ranks within each row and each layer of the classification. So far as we know, neither of these generalizations has been carried out.

Finally, it does not seem possible to obtain a distribution-free test for interactions by this method.

Median tests

37.44 A different approach to the construction of distribution-free AV tests was followed by Brown and Mood (1951). The principle of test construction which they used is (a) to estimate all parameters unspecified by the hypothesis by median statistics; and then (b) to test whether the residuals from this median-estimated model have half of their signs negative and half positive.

For example, in the one-way classification, with n_j observations in the j th group, $\sum_{j=1}^k n_j = n$, only the general mean μ is left unspecified by the hypothesis of equal group means. We therefore estimate μ by the median \tilde{y} of the n (assumed even) observations. In each group, we now see how many observations lie below and above \tilde{y} . We obtain the table:

Group:	1	2	k	TOTAL
No. of observations $\begin{cases} \geq \tilde{y} \\ < \tilde{y} \end{cases}$	m_1 $n_1 - m_1$	m_2 $n_2 - m_2$	m_k $n_k - m_k$	$\frac{1}{2}n$ $\frac{1}{2}n$
TOTAL	n_1	n_2	n_k	n

(37.76)

The hypothesis now is that all k groups have the same median, i.e. $E(m_j) = \frac{1}{2}n_j$ for each j . It will be seen at once that this is the binomial homogeneity test treated in 33.55. The statistic (33.122), which in the present notation is

$$X^2 = \sum_{j=1}^k \frac{(m_j - \frac{1}{2}n_j)^2}{\frac{1}{4}n_j}, \quad (37.77)$$

is asymptotically distributed in the χ^2 form with $(k-1)$ d.fr. The test (which may be carried out exactly by the method of 33.19, Case 1) is distribution-free. As for the Sign test in 32.3, the use of the median reduces the problem to a binomial one.

37.45 More complex AV situations may be treated by the same general method. For the two-way cross-classification, as Exercises 37.18–20 indicate, a variety of tests is available. With one observation per cell, or in the more general situation when there are no interactions, column- (or row-) effects may be tested; when interactions are present, column-effects may be tested against interactions, or column- and interaction-effects may be tested jointly.

37.46 These median tests are attractive because of their computational simplicity and the fact that their theory is immediately available, at least in large samples, from that of $(2 \times c)$ contingency tables. However, not every problem is soluble by their use, e.g. there is no test known for column-effects against residual in the general balanced two-way cross-classification. Further, even when a test is available, it is not always distribution-free—Brown and Mood (1951) show that a median test for interactions in the balanced two-way classification is not. Finally, the efficiency of these median tests is not generally as high as that of tests using ranks or normal scores when the errors are near-normal. In 32.6–7, the Sign test was found to have ARE of $2/\pi$ in the normal case, against $3/\pi$ for the ranks test; and in Exercise 31.12, a median test of randomness was seen to have ARE of 0.78 against the 0.98 found in 31.38 for tests using ranks. Andrews (1954) showed that the one-way classification test discussed in 37.44 has the same ARE, $2/\pi$, as the Sign test while, as we saw in 31.71, the comparable test statistic (31.150) based on ranks has ARE $3/\pi$.

Bhapkar (1963) gave some efficiency results for the two-way classification median test.

37.47 The restricted scope and relatively low efficiency of the median tests obtained by using the principle given in 37.44 are rather disappointing—intuitively, it seems that it ought to be possible to find general AV procedures with the high efficiencies

which we saw in Chapter 31 to be characteristic of tests based on ranks. The nearest approaches to such procedures have been developed in a series of papers by Lehmann (1963a, b, c, 1964) and by Hodges and Lehmann (1963) (cf. also Høyland (1965) and Bickel (1965)). These procedures are only asymptotically distribution-free, and it is interesting that they, too, are based on median estimation methods of a different sort from those in 37.44.

37.48 Suppose that

$$x_{ip} = \mu_i + \varepsilon_{ip} \quad (i = 1, 2, \dots, c; p = 1, 2, \dots, n_i)$$

is the model for a set of $n = \sum_{i=1}^c n_i$ observations. The ε_{ip} are independent, but otherwise we assume only that they have the same distribution. We write

$$\theta_{ij} = \mu_i - \mu_j$$

for the parameters in terms of which all quantities of interest may be expressed. We shall discuss median estimators $\tilde{\theta}_{ij}$ of θ_{ij} .

Let y_{ij} be the median of the $n_i n_j$ differences $(x_{ip} - x_{jq})$, where $p = 1, 2, \dots, n_i$; $q = 1, 2, \dots, n_j$. The y_{ij} are clearly estimators of the θ_{ij} , but they do not possess the desirable transitivity property that

$$\tilde{\theta}_{ij} + \tilde{\theta}_{jk} + \tilde{\theta}_{ki} = 0 \quad \text{for all } i, j, k. \quad (37.78)$$

Adjusted estimators which satisfy (37.78) are

$$\tilde{\theta}_{ij} = y_i - y_j, \quad (37.79)$$

where $y_i = \frac{1}{c} \sum_{l=1}^c y_{il}$. Lehmann (1963a) gives a numerical illustration showing that the $\tilde{\theta}_{ij}$ agree well with the usual LS estimators $\hat{\theta}_{ij}$.

As $n \rightarrow \infty$ suitably, the $\tilde{\theta}_{ij}$ tend to multivariate normality; they also have the same estimation efficiency, compared with the standard AV estimators based on means, as the Wilcoxon test has compared to "Student's" t -test. If f is the common frequency function of the ε_{ip} , it follows from (31.115) that the efficiency is

$$12\sigma^2 \left[\int_{-\infty}^{\infty} \{f(x)\}^2 dx \right]^2 = k^2, \quad (37.80)$$

say. By 31.60-1, k^2 may be infinite, but can never be less than 0.864 for any continuous f ; in the normal case, $k^2 = \frac{3}{\pi} \doteq 0.95$. Thus, generally, $kn^{\frac{1}{2}}(\tilde{\theta}_{ij} - \theta_{ij})$ has the same limiting distribution as $n^{\frac{1}{2}}(\hat{\theta}_{ij} - \theta_{ij})$, where $\hat{\theta}_{ij}$ is the standard (LS) AV estimator of θ_{ij} .

37.49 The asymptotic property stated in the last sentence of 37.48 implies that, provided that we can estimate k^2 at (37.80) consistently, we may set up analogues of all the usual AV procedures in terms of the median estimators $\tilde{\theta}_{ij}$. Lehmann (1963c) gives two consistent estimators of k^2 and (1963b) develops large-sample confidence intervals for any contrast or set of contrasts in the parameters. Further, the same author (1964) (see also Hodges and Lehmann (1962)) extends his results to the situation

where there are "nuisance factors" in the observations (i.e. in the terminology of Chapter 38, "blocks" within the experiment) with equal numbers of observations in each cell of the same "block."

Bhuchongkul and Puri (1965) extend the asymptotic theory to a class of estimators of contrasts including those based on normal scores.

Missing observations

37.50 The advantages of balanced arrangements in Model I AV, namely orthogonality, ease of computation and superior robustness, are such that most designed analyses will seek to take advantage of them. Nevertheless, force of circumstances will sometimes lead to involuntary departures from the intended equality of frequencies: plants or animals may die, human subjects may prove reluctant to co-operate, or records may be lost before analysis. If this happens, we are always free to analyse the achieved unequal frequencies by the appropriate non-orthogonal methods, but, as we have seen, these are often complicated. Moreover, accidental losses of observations are rarely extreme; usually only one or a few are found to be missing. It is therefore worth investigating whether we can retain the original AV structure and correct it for the missing observations, rather than abandon it altogether.

37.51 Suppose, then, that m of the n intended observations are missing. Without loss of generality, we take these to be the last m components of the observation vector \mathbf{y} , which now become unknowns, say u_1, \dots, u_m . Thus we may write $\mathbf{y} = \begin{pmatrix} \mathbf{z} \\ \mathbf{u} \end{pmatrix}$ where \mathbf{z} $((n-m) \times 1)$ contains the actually observed values of y and \mathbf{u} $(m \times 1)$ the unknown observations. In effect, we are presented with a fresh set of unknowns to estimate, in addition to the original parameters of the model. It is natural, in these circumstances, to estimate the values in \mathbf{u} by the same LS method as we use for the original parameters.

37.52 The sum of squared residuals

$$S = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$

must therefore now be minimized not only for variation in $\boldsymbol{\theta}$ (as was done in 19.4) but also for variation in \mathbf{u} . If we first minimize S with respect to $\boldsymbol{\theta}$, we shall, of course, obtain the original LS solution (i.e. the LS solution if there had been no missing observation), but the estimator and the Residual SS of that solution will now both be functions of \mathbf{u} , say $\hat{\boldsymbol{\theta}}(\mathbf{u})$ and $S_0(\mathbf{u})$. The minimization process could now be completed by minimizing $S_0(\mathbf{u})$ for variation in \mathbf{u} .

However, this two-stage minimization procedure, which was suggested by Yates (1933), is not the easiest way in general. Instead, let us minimize S first for variation in \mathbf{u} . Partitioning \mathbf{X} into $\begin{pmatrix} \mathbf{X}_z \\ \mathbf{X}_u \end{pmatrix}$ conformably with the partition $\mathbf{y} = \begin{pmatrix} \mathbf{z} \\ \mathbf{u} \end{pmatrix}$, we have

$$S = (\mathbf{z} - \mathbf{X}_z\boldsymbol{\theta})'(\mathbf{z} - \mathbf{X}_z\boldsymbol{\theta}) + (\mathbf{u} - \mathbf{X}_u\boldsymbol{\theta})'(\mathbf{u} - \mathbf{X}_u\boldsymbol{\theta}). \quad (37.81)$$

Since only the second of the two non-negative terms on the right of (37.81) depends upon \mathbf{u} , we reduce it to zero by putting

$$\mathbf{u} = \mathbf{X}_u\boldsymbol{\theta}; \quad (37.82)$$

thus S at (37.81) is reduced to its first term, which may then be minimized with respect to θ .

But the results of the two two-stage minimization methods just described must be the same. Thus if we obtain $\hat{\theta}(u)$ by the first method, which gives

$$\hat{\theta}(u) = (X'X)^{-1}X'y = (X'X)^{-1}(X'_z z + X'_u u), \quad (37.83)$$

and use this in conjunction with (37.82), we have

$$u = X_u \hat{\theta}(u). \quad (37.84)$$

(37.84) states that each missing observation is to be equated to its estimated expectation in the original LS analysis.

37.53 (37.84) is a set of linear equations to be solved for u , and the solution \hat{u} is then to be used in the original LS analysis. A straightforward solution was given by Tocher (1952).

First, suppose that we replace u by the null vector 0 in the original LS analysis. (37.83) becomes

$$\hat{\theta}(0) = (X'X)^{-1}X'_z z. \quad (37.85)$$

Now consider

$$\hat{u} = \{I - X_u (X'X)^{-1} X'_u\}^{-1} X_u \hat{\theta}(0). \quad (37.86)$$

Observing from (37.83) that

$$\hat{\theta}(u) = \hat{\theta}(0) + (X'X)^{-1} X'_u u,$$

we find that (37.86) reduces to

$$\{I - X_u (X'X)^{-1} X'_u\} \hat{u} = X_u \hat{\theta}(0) - X_u (X'X)^{-1} X'_u \hat{u},$$

so that

$$\hat{u} = X_u \hat{\theta}(\hat{u}). \quad (37.87)$$

Thus \hat{u} defined by (37.86) is the solution of (37.84).

37.54 We have seen, therefore, that in order to estimate the m missing observations u , so that we may preserve the computational form of the original LS analysis, we need only

- (a) perform the original analysis with $u = 0$ to obtain $\hat{\theta}(0)$ at (37.85);
- (b) calculate \hat{u} at (37.86); and
- (c) again perform the original analysis using \hat{u} in y .

It should be noted that the matrix in braces to be inverted in (37.86) is $(m \times m)$. Thus, if only one observation is missing, the matrix is a scalar and stage (b) above is very simple.

The second of the four papers by Wilkinson (1957-60) on missing observations gives detailed solutions of (37.84) for many common AV situations. See also Biggers (1959).

37.55 It is easy to see that the estimator $\hat{\theta}(\hat{u})$ obtained by using (37.86) in the original LS analysis is exactly the estimator which would have been obtained by using

the $(n-m)$ observed values alone (generally in a non-orthogonal analysis). For, from (37.83),

$$\begin{aligned}\mathbf{X}'\mathbf{X}\hat{\mathbf{u}} &= \mathbf{X}'_z\mathbf{z} + \mathbf{X}'_u\hat{\mathbf{u}} \\ &= \mathbf{X}'_z\mathbf{z} + \mathbf{X}'_u\mathbf{X}_u\hat{\mathbf{u}}\end{aligned}$$

using (37.87). Thus

$$\begin{aligned}\mathbf{X}'_z\mathbf{z} &= (\mathbf{X}'\mathbf{X} - \mathbf{X}'_u\mathbf{X}_u)\hat{\mathbf{u}} \\ &= \mathbf{X}'_z\mathbf{X}_z\hat{\mathbf{u}}\end{aligned}\tag{37.88}$$

and (37.88) is precisely the set of equations satisfied by $\hat{\mathbf{u}}$ when \mathbf{z} alone is analysed. This result, together with the disappearance of the second term on the right of (37.81), implies at once that the Residual SS obtained by using $\hat{\mathbf{u}}$ in the original LS analysis is identical with that obtained when \mathbf{z} alone is analysed. However, the degrees of freedom for the Residual SS must obviously be reduced, since we now have only $(n-m)$ observations. If \mathbf{X} and \mathbf{X}_z both have the same rank (e.g. when both have full rank) the Residual SS will have its d.fr. reduced by m , the number of missing observations. More generally, the reduction will be (cf. Exercise 19.8) m minus the difference in rank between \mathbf{X} and \mathbf{X}_z .

37.56 Although the Residual SS requires no adjustment, all the other SS in the AV table will be incorrect if $\hat{\mathbf{u}}$ at (37.86) is used in the original LS analysis. This is most easily seen from the fact (cf. Example 35.4, 35.38 and Example 35.6) that each other SS in the AV table may be obtained as the difference between the Residual SS in two linear models, one of which is a restricted form of the other. Evidently, any of these Residual SS is correctly obtainable, by the argument of 37.50-5, by using $\hat{\mathbf{u}}$ at (37.86) for that model, but of course $\hat{\mathbf{u}}$ will in general differ from that in the full model considered so far. Thus each of these Residual SS will be too large if $\hat{\mathbf{u}}$ for the full model is used, since it will not be the correct (minimizing) $\hat{\mathbf{u}}$. Hence the other SS in the AV table need correction by the difference between the subtractive corrections to the corresponding Residual SS (or by a single subtractive correction if one of the latter is the Residual SS for the full model). Correspondingly, degrees of freedom must be corrected by the difference between two adjustments of the form discussed in 37.55, but this difference will often be zero.

In the third of his four papers, Wilkinson (1957-60) gives an explicit method of obtaining the subtractive corrections to the other Residual SS, and hence the other SS in the AV table. Fortunately, as Yates (1933) pointed out, the latter corrections, being generally differences of quantities of the same sign, are often small, and the unadjusted SS may be used as approximations.

37.57 Tocher (1952) gives similar methods of analysis for other types of "spoilt" experiments, namely those in which some observations are irretrievably mixed up and those in which some observations are unwittingly duplicated—Plackett (1950) also discusses the latter situation.

EXERCISES

37.1 There are G groups of observations, and all observations within a group are normally distributed with common mean and common variance σ_g^2 , the model (37.1) holding except for

the homoscedasticity condition below it. Consider the sets of constraints

C_1 : all the σ_v^2 are equal ($G-1$ constraints);

C_2 : r of the k parameters in θ are zero.

Working in terms of the variable $z_\lambda = y_\lambda J^{-\frac{1}{n}}$, so that (37.3) reduces to

$$L_\lambda(z | \hat{\theta}, \hat{\sigma}^2) = \{\hat{\sigma}_\lambda^2(z)\}^{-1n},$$

show that (37.8) gives

$$L(z | \hat{\lambda}_{(2)}) = L(z | \hat{\lambda}) l_1(z) l_2(z)$$

where l_1 is the LR test statistic defined at (24.40), and $l_2 = \left\{1 + \frac{r}{n-k} F\right\}^{-1n}$, where F is the variance-ratio test statistic defined generally at (24.99) and for this case in Example 24.8. (This result generalizes Exercise 24.6.)

(Box and Cox, 1964)

37.2 Using Exercise 23.7, show that if in (37.8) l_p is distributed free of certain parameters for which there is a complete sufficient (vector) statistic t , and l_q is a function of t alone, l_p and l_q are stochastically independent. Apply this result to establish the independence results in Exercises 24.6 and 24.13. Show in Exercise 37.1 that $l_1(z)$ and $l_2(z)$ are independent when C_1 and C_2 both hold.

(Cf. Hogg, 1961)

37.3 In fitting orthogonal polynomials of degree k as in 28.16, the reduction in the total SS associated with the term of degree r is $Q_r = \hat{\alpha}_r^2 \sum_{j=1}^n \phi_j^2(x_j)$ as at (28.72), Vol. 2. Show that the ratios

$$z_r = Q_{k-r+1} / \sum_{s=k-r+2}^{k+1} Q_s, \quad r = 1, 2, \dots, k,$$

where $Q_{k+1} = (n-k)s^2$ is the Residual SS, are all independently distributed when the regression coefficients α_r are all zero:

- (a) by using the result of Exercise 37.2; and
- (b) by using the result of Exercise 23.27.

(This result indicates (cf. Hogg (1961)) that one may independently test the regression coefficients if one starts from the highest order and works downwards, "pooling" the associated SS of those adjudged zero with the Residual SS, until one is adjudged non-zero, when the process stops. All the tests are, of course, t^2 (F) tests, and the overall test has size $1 - (1 - \alpha)^k \sim k\alpha$ if a test of size α is used at each stage. T. W. Anderson (1962) shows under weak assumptions that this procedure maximizes the probability of correctly locating a non-zero coefficient.)

37.4 In 37.10, show that for the binomial distribution of 5.4, (37.14) gives the variance-stabilizing transformation $u\left(\frac{x}{n}\right) = \arcsin \left\{ \left(\frac{x}{n}\right)^{\frac{1}{2}} \right\}$, where x/n is the observed proportion of "successes."

(Anscombe (1948) shows that better variance stabilization is obtained if x/n is replaced by $(x + \frac{3}{8})/(n + \frac{3}{4})$. Freeman and Tukey (1950)

suggest $\arcsin \left\{ \left(\frac{x}{n+1}\right)^{\frac{1}{2}} \right\} + \arcsin \left\{ \left(\frac{x+1}{n+1}\right)^{\frac{1}{2}} \right\}$

(Cf. Example 37.1).)

37.5 In 37.10, show that for the negative binomial distribution of 5.15, (37.14) gives the

variance-stabilizing transformation $u(t) = \arcsinh \left\{ \left(\frac{x}{n} \right)^{\frac{1}{2}} \right\}$ where x/n is the observed proportion of "successes."

(Anscombe (1948) shows that better variance stabilization occurs if x/n is replaced by $(x + \frac{3}{8})/(n - \frac{3}{4})$.)

37.6 In Exercise 37.4, show that the alternative transformation $u\left(\frac{x}{n}\right) = \log \left\{ \frac{x}{n} / \left(1 - \frac{x}{n}\right) \right\}$ stabilizes the variance near $p = \frac{1}{2}$. Show that this transformation is strictly appropriate when (37.9) is $D_n^2(\theta) = c\theta^2(1-\theta)^2$.

(Cf. Bartlett, 1947a)

37.7 For a cross-classification with unequal cell frequencies, show that if the cell means are analysed as single observations, their average variance may be estimated by s^2/H , where s^2 is the Residual MS of the original observations and H is the harmonic mean of the cell frequencies. Hence show how an approximate AV of the cell means may be carried out.

(Cf. Scheffé, 1959)

37.8 Applying the method of Exercise 37.7 to the numerical data of Exercise 35.7, show that the approximate AV for cell means is

	SS	d.fr.	MS
Between sexes	0.023	1	0.023
Between breeds	0.020	7	0.003
Interactions	0.006	7	0.001
Residual	0.049	15	0.0017 (= 0.0227 × 0.0759)
		517	

Compare the values of the F -ratios in this table with the exact values in Exercise 35.7.

37.9 Show that if a linear model contains terms $\theta_i x_i^{\mu_i}$, $\mu_i \neq 0$, we have approximately

$$x_i^{\mu_i} = x_i + (\mu_i - 1)x_i \log x_i.$$

Hence show how μ_i can be estimated. The process may be iterated.

(Box and Tidwell, 1962)

37.10 In 37.28, show that $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is invariant under any non-singular transformation $\mathbf{W} = \mathbf{X}\mathbf{T}$. Hence, taking $\mathbf{W}'\mathbf{W}$ to be diagonal, show that $m = \sum_r M_{rr}^2$ satisfies

$$m = \frac{(N-2)^{(2)}}{(N+1)^{(3)}} \left\{ \sum_{i=1}^p \left(\frac{k_{ii}}{k_{ii}^2} \right) + \sum_{i \neq j=1}^p \sum \left(\frac{k_{22}}{k_{20} k_{02}} \right)_{ij} \right\} + \frac{(N-1)p(p+2)}{N(N+1)}.$$

(Box and Watson, 1962)

37.11 In 37.31, by considering the variance of the diagonal elements M_{rr} of \mathbf{M} , show that $m = \sum_r M_{rr}^2 \geq p^2/N$. Using the invariance property of Exercise 37.10, suppose that $\mathbf{X}'\mathbf{X} = \mathbf{I}$, and adjoin $(N-p)$ further columns, one of which is $N^{-1}\mathbf{1}$, to \mathbf{X} to form an orthogonal matrix.

Hence show that $m \leq p(N-1)/N$. Show from these derivations that the lower bound to m , but not the upper, can be attained. (Box and Watson, 1962)

37.12 In 37.32, show that if we choose $(p+1)$ group-frequencies n_j in the one-way classification so that

$$\sum_{j=1}^{p+1} \frac{1}{n_j} = \frac{(N-1)}{N(N+1)} p^2 + \frac{4}{N+1} p + \frac{1}{N} \sim \frac{p^2 + 4p + 1}{N+1},$$

C_X at (37.52) is zero and there is no correction to $V_P(W)$ for non-normality, whatever the underlying distribution of the errors. If $p = 1$, and $n_1 = rN$, $n_2 = (1-r)N$, show that $C_X = 0$ when

$$r = \frac{1}{2} \left\{ 1 \pm \left(\frac{N-2}{3N} \right)^{\frac{1}{2}} \right\} \sim \frac{1}{2} (1 \pm 3^{-1/2}),$$

and that if $N = 12$, the optimal integer group-frequencies are 9 and 3. (Box and Watson, 1962)

37.13 Establish (37.71-2) by writing U defined below (37.70) in the form

$$U = \sum_{i \neq l} \sum R_{il}$$

where

$$R_{il} = \sum_{j=1}^c y_{ij} y_{lj},$$

and showing that

$$E_P(R_{il}) = 0, \quad E_P(R_{il}^2) = (c-1)(k_2)_i(k_2)_l,$$

and hence

$$E_P(U) = 0, \quad E_P(U^2) = (c-1) \sum_{i \neq l} (k_2)_i(k_2)_l.$$

(Pitman, 1938)

37.14 In 37.41 show that when (37.75) holds,

$$V_P(W) = \frac{2(r-1)}{r^3(c-1)},$$

and hence by the method of 37.29 that the d.fr. of the approximate F -test should be adjusted to

$$\nu_1 = (c-1) - \frac{2}{r},$$

$$\nu_2 = (r-1)\nu_1.$$

Show that when the ranks $1, 2, \dots, c$ in each row are used as conventional numbers, with R_j as the sum of the ranks in the j th column and $T = \sum_{j=1}^c \left\{ R_j - \frac{r(c+1)}{2} \right\}^2$, the statistic W reduces to

$$W = \frac{12T}{r^2 c(c^2-1)},$$

and that as $r \rightarrow \infty$,

$$\nu_2 W / (1-W) \sim \frac{12T}{rc(c+1)}$$

has a χ^2 distribution with $(c-1)$ d.fr.

(Friedman (1937); Kendall and Babington Smith (1939); Friedman (1940) compares the F and χ^2 approximations.)

37.15 In Example 37.1, expand

$$u_c(t) = (\theta + c)^{\frac{1}{2}} \left(1 + \frac{t - \theta}{\theta + c} \right)^{\frac{1}{2}}$$

in series and show that

$$E(u_c) = (\theta + c)^{\frac{1}{2}} - \frac{1}{8}\theta^{-\frac{1}{2}} + \frac{24c - 7}{128}\theta^{-\frac{3}{2}} + o(\theta^{-\frac{3}{2}}),$$

and

$$\text{var } u_c = \frac{1}{4} \left\{ 1 + \frac{3 - 8c}{8\theta} + \frac{32c^2 - 52c + 17}{32\theta^2} + o(\theta^{-2}) \right\},$$

so that the choice $c = \frac{3}{8}$ removes the term of order θ^{-1} in the variance, reducing it to

$$\text{var } u_{3/8} \sim \frac{1}{4} \left(1 + \frac{1}{16\theta^2} \right).$$

Hence show that

$$\{E(u_c)\}^2 - c \sim \theta - \frac{1}{4} - \frac{3 - 8c}{32\theta}$$

so that if the inverse transformation is used on u_c to obtain an estimator of θ , its downward bias is nearly constant at $\frac{1}{4}$.

(The $c = \frac{3}{8}$ result is due to A. H. L. Johnson; cf. Anscombe (1948).)

37.16 In Exercise 37.15, show that the coefficients of skewness and kurtosis of $u_c(t)$ are

$$\gamma_1 = -\frac{1}{2\theta^{\frac{1}{2}}} \left\{ 1 + \frac{25 - 48c}{16\theta} \right\} + o(\theta^{-\frac{3}{2}}),$$

$$\gamma_2 = \frac{1}{\theta} \left\{ 1 + \frac{945 - 1536c}{256\theta} \right\} + o(\theta^{-2}),$$

compared with

$$\gamma_1 = \theta^{-\frac{1}{2}},$$

$$\gamma_2 = \theta^{-1},$$

for the original Poisson variable t . Thus, whatever c is chosen, γ_1 is approximately halved (with changed sign) and γ_2 is unaffected to the first order.

(Anscombe, 1948.)

37.17 Using the result for $\text{var } u_c$ in Exercise 37.15, show that the transformation

$$u'_\delta = (t + \frac{1}{2} + \delta)^{\frac{1}{2}} + (t + \frac{1}{2} - \delta)^{\frac{1}{2}}$$

has variance

$$\text{var } u'_\delta = 1 - \frac{1}{8\theta} + \frac{16\delta^2 - 1}{32\theta^2} + o(\theta^{-2}),$$

so that if we choose $\delta = \frac{1}{2}$ to give u' of Example 37.1,

$$\text{var } u'_\frac{1}{2} = 1 - \frac{1}{8\theta} + \frac{3}{32\theta^2} + o(\theta^{-2}).$$

37.18 For the two-way cross-classification with one observation per cell, or with more than one if there are no interactions, show that a median test for column-effects is obtained by counting the number m_j of observations in the j th column which exceed the row median, and forming a $(2 \times c)$ table like (37.76), with (37.77) as the large-sample test statistic.

(Brown and Mood, 1951)

37.19 For the general balanced two-way cross-classification, show that column-effects may be tested against interactions by finding the median in each cell and applying the test of Exercise

37.18 to these medians. Show that the test remains distribution-free if cell frequencies differ between rows, but not within rows. (Brown and Mood, 1951)

37.20 In Exercise 37.19, show that column-effects and interactions may be tested jointly by counting the number m_{ij} of the n_{ij} observations in the (i, j) th cell which exceed the i th row median and then testing that $E(m_{ij}) = \frac{1}{2}n_{ij}$ for all i, j . Show that this is equivalent to the hypothesis (cf. 33.60) that in each row of a $(r \times c \times 2)$ three-way table, the columns and layers are independent, leading to a large sample χ^2 test with $r(c-1)$ d.f. (Brown and Mood, 1951)

CHAPTER 38

THE DESIGN OF EXPERIMENTS

38.1 For the greater part of this work, we have been concerned with the problems arising in the analysis of observations, principally the problems of estimation and testing which appear in various theoretical contexts. In a very obvious way, every investigation of a method of analysis carries its own lesson for the future. Thus, e.g., when we learn that a particular method of estimation is more efficient than another, the immediate implication is that we should use the better method in future. However, this implication alone would not lead us to modify the *method* of making observations in future, but merely to modify the *analysis* of the observations. In this chapter and the two following, we shall be discussing questions of design, by which we mean considerations affecting the method of making, or collecting, the observations to be analysed.

38.2 Design considerations are not entirely new to us. In Example 28.4 we discovered from the analysis of a simple linear regression problem that by choosing the origin of measurement and the values of the regressor in a certain way, we could ensure an orthogonal analysis and also minimize the sampling variances of our estimators. This is a design question, because it relates to how the observations are to be made. In the same Example, we remarked the hazard that this optimum solution removes the possibility of checking the assumption that the regression model was linear in the value of the regressor. Unless we were very sure on this point, we should probably "hedge" slightly by departing from the optimum choice of regressor values.

Again, in 37.21, the results of robustness studies implied that equal frequencies should be used in all cells of an experiment. Once more, this is a design question since it affects the method of making observations.

38.3 In this chapter we shall discuss questions of design as they affect experimentation, largely using the linear models and AV techniques of Chapters 35–7. In Chapters 39–40, we shall turn to design problems in sample surveys. The distinction between these fields is fairly clear-cut, and may be expressed by saying that in surveys we make observations on a sample taken from a finite population of individuals, whereas in experiments we make observations which are in principle generated by a hypothetical infinite population, in exactly the way that the tosses of a coin are (cf. 1.29 and 9.4, and Example 38.1 below). Of course, we may sometimes experiment on the members of a sample resulting from a survey, or even make a sample survey of the results of an (extensive) experiment, but the essential distinction between the two fields should be clear.

Cochran (1965) gives an interesting general discussion of inferential problems which arise particularly in surveys rather than in controlled experiments.

Principles of experimentation

38.4 Classical discussions of the principles of experimentation emphasized the importance of varying the (supposedly) causal factors in an experiment in order to observe the effect upon the dependent variable being studied. In two respects, however, these discussions are now generally seen to have been inadequate. Firstly, they tended to be phrased in terms of the variation of a single causal factor at a time, rather than of all the causal factors in combination. Thus, J. S. Mill's (1843) fifth canon of experimental enquiry states:

"Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation."

In the light of the results of Chapter 35, we see now that a "one-at-a-time" approach can have no hope of evaluating the interactions between causal factors. Not only does this deprive us of essential knowledge of the linkages between causal factors: it may actually be positively misleading. For suppose that it is the purpose of an experiment to find which combination of ingredient *A* and ingredient *B* gives the highest resistance to breakage in a ceramic product. If we find the dose of ingredient *A* which gives highest resistance, and the dose of *B* which does so, it is by no means true that if we combine these values we have arrived at the optimum combination sought, as the reader may easily convince himself numerically. Interaction between the factors, which can produce effects like this, can only be studied by varying them simultaneously.

38.5 The second inadequacy of the classical discussions is even more radical, and is again illustrated by the quotation from J. S. Mill in 38.4. It arises from the danger of attributing to one or more of the experimental factors, effects upon the dependent variable which are in reality due to variations in some causal factors not included in the experiment. An unrecognized causal factor may (unknown to the experimenter) vary during the course of the experiment in such a way as to favour a particular combination of experimental factors; this combination will then appear to be highly effective, when it is really the unrecognized factor which is producing the good results.

The classical discussions had no solution to this problem, and it is essential to realize how deep-seated and ever-present the problem is. We can *never* be quite sure that all the important, or even the most important, causal factors have been incorporated in the structure of the experiment. Some may be quite unknown; others, although known, may wrongly be considered to be of minor importance and deliberately neglected. We always need to guard against the perversion of the inferences within an experiment by adventitious outside effects.

Randomization

38.6 The modern solution to the problem of 38.5 was first propounded by R. A. Fisher in the nineteen-twenties—cf. especially his book (Fisher (1935)). We have seen throughout this work that his contributions to statistical theory were remarkable and far-ranging. Nevertheless, it is probably no exaggeration to say that his advocacy

of randomization in experiment design was the most important and the most influential of his many achievements in statistics.

38.7 The principle of randomization is simply stated: *Whenever experimental units are allocated to factor-combinations in an experiment, this should be done by a random process using equal probabilities.* Thus every factor-combination will have the same chance of being applied to each eligible experimental unit.

Evidently, even if we randomize in accordance with this principle, the particular allocation of experimental units which we make may still work in favour of particular factor-combinations. However, the difficulty of 38.5 no longer troubles us if we incorporate the process of randomization into the framework within which our inferences are made. The hypothetical population within which we now infer includes every possible pattern of allocation of experimental units which the randomization could have produced. Within this population, by the very nature of the randomization process, the effects of factors outside the experiment can show no favour to the factors inside it, and our inferences are free from bias.

Even if the relationship of the dependent variable with some unsuspected causal factor is not recognized until after the experiment, the validity of the inferences will not be impaired, provided that that factor's influence was "randomized out" of the experiment.

38.8 Thus, the problem of 38.5 is solved by changing the inferential base. Necessarily, this has the effect of changing the theoretical basis of our inference, and we shall develop this point shortly. First, we illustrate by a simple example.

Example 38.1

An experiment is to investigate the dependence of reaction-time in male automobile drivers upon the alcohol content of their blood. The drivers taking part are to consume measured doses of alcohol and, after a fixed time-lapse, to undergo a blood-alcohol test and certain standardized tests of reaction-times. The problem is how the drivers are to be allocated to the different alcohol-doses.

This is intrinsically a regression problem, with reaction-time as dependent variable (y) and blood-alcohol content as regressor (x), but it should be observed that x is not strictly under control—we can only control the alcohol-dose (z), and we merely observe the value of x in each case. However, z and x will be in fairly close relationship, and it is reasonable to assume that to each fixed value of z , there will correspond a grouped set of values of x . If the z -values are sufficiently well spaced, the x -groups will not overlap, and we can treat the problem as a one-way classification, the classification being indexed by the values of z .

In this example, it is not difficult to see the problems which arise in the absence of a randomized allocation of alcohol-doses to drivers. Suppose, for example, that the drivers were allowed themselves to choose their doses of alcohol. Presumably, hard drinkers would choose larger doses than other drivers. Since normal drinking habits affect one's tolerance of alcohol, the results in the reaction-time tests might tend to mask the true differences between the effects of the various alcohol-doses.

It may be argued that allowing the drivers to choose their doses in fact gives a truer picture of what happens in driving practice. Even if this be true, it is important to realize that it is irrelevant to the scientific enquiry undertaken. The essential difference between a survey and an experiment is that the former attempts to delineate an existing population, while the latter is concerned to investigate relationships which need not be those in any precise population—this was the distinction in 38.3 above. We shall return to this point in 38.12; for the present example, it should be clear that a randomized allocation of alcohol-doses to drivers is needed to protect the inferences within the experiment from bias.

38.9 The fact that an "outside" influence, like normal drinking habits in Example 38.1, has been removed by randomization in no way precludes us from analysing its effect after the (randomized) experiment is complete. In the case of Example 38.1, it would presumably be a difficult matter to ascertain at all accurately the normal drinking habits of the participants. However, consider the effect of the times of day at which the tests were taken. Whether or not this has been "randomized out" of the experiment, there is nothing to prevent our subsequently carrying out a regression analysis of reaction-time upon time of day. If we found, say, that tests taken later in the day tended to have higher reaction-times, this would be a matter for investigation and might lead to a modification of the experimental procedure on future occasions.

38.10 Our statement of the virtues of randomization must not be taken to imply that all randomized experiments leave nothing to be desired. Consider again the effect of time of day upon reaction-times, as in 38.9, and suppose that all the tests in the experiment were carried out at 6 p.m., the end of the day's work of the drivers taking part. In effect, the factor "time of day" is then constant at one level, and the experiment is vulnerable to the possibility that this factor interacts with blood alcohol-content in its effect upon reaction-times. The randomization with respect to alcohol-doses does not help at all in this respect, and the criticism of classical procedure in 38.5 applies here. In fact, the randomization has been incomplete, because time of day has been neglected as a possible causal factor. Randomization can only confer inferential benefits within the sphere to which it has been applied.

38.11 It will be clear, then, that the factors influencing the dependent variable in any experiment are, explicitly or implicitly, divided by the experimenter into three classes:

- (1) those incorporated into the structure of the experiment (alcohol-dose in Example 38.1);
- (2) those "randomized out" of the experiment (normal drinking habits in Example 38.1); and
- (3) those neither incorporated nor randomized out (time of day in 38.10).

It will be observed that classes (1) and (2) require positive action, affecting the actual layout of the experiment or of the randomization procedure employed. By contrast, the last of our three classes is a residual one. It is true that the experimenter

may deliberately decide that a certain factor is negligible, so that it does not even require randomization to remove its possible effects—in Example 38.1, the eye-colour of the driver is almost certainly negligible in this way. However, a factor may find its way into class (3) simply from being overlooked, like time of day in 38.10.

A substantial part of the skill of the experimenter lies in his choice of factors to be randomized out of the experiment. If he is careful, he will randomize out all the factors which are suspected to be causally important but which are not actually part of the experimental structure. But every experimenter necessarily neglects some conceivably causal factors; if this were not so, the randomization procedure required would be impossibly complicated. Thus the choice of factors to be randomized out is essentially a matter of judgement.

38.12 We saw in 38.7–8 that the population, within which inferences from a randomized experiment may validly be made, is a hypothetical one depending upon the process of randomization itself. The experimenter, however, must apply his inferences to the real world. In Example 38.1, the hypothetical population for the inferences drawn from the experiment includes every possible allocation of alcohol-doses to the drivers taking part. How far could these inferences be extended to cover the larger populations of

- (a) all male automobile drivers;
- (b) all automobile drivers, females included?

If the drivers taking part in the experiment were a random sample (not necessarily *simple* random) of all male drivers, few experimenters would hesitate to generalize the findings of the experiment to population (a). Similarly, few experimenters would be rash enough to generalize to population (b) without further knowledge, perhaps from other experiments on female drivers. However, suppose that the drivers taking part were selected from the employees of one corporation. Even if they were randomly so selected, this would only give us comfort in generalizing to the limited population of all drivers employed by that corporation. If (as is commonly the case) the corporation was not chosen by any other process than its own self-interest or its willingness to co-operate with some scientific body, further generalization of the results of the experiment is a matter of judgement. Only in so far as the experimental material is, or is judged to be equivalent to, a random sample from a larger population may we generalize the experimental results to that population.

There is ultimately no escape from the use of judgement in this connexion, for there are always the problems of generalization in space (e.g. to other countries) and in time.

38.13 We have dealt in a very compressed way with some of the fundamental questions of experimental inference. For a fuller (and largely non-technical) discussion, the reader is recommended to read the book by D. R. Cox (1958a), as well as that of Fisher (1935).

Nuisance factors: block experiments

38.14 We have seen that an essential feature of modern experimentation is the "randomization out" of the experiment of the effects of factors outside the experimental structure. In practice, such effects are often produced by the fact that the experimental units themselves cannot be physically identical. Thus, agricultural experiments are carried out on plots of land which (however close together) will not have identical fertility characteristics. What is more, as the number of plots required for the experiment increases, the variation in their fertility will probably also increase. Thus it appears advisable to use a number of small groups of similar plots, rather than a single large group of rather heterogeneous plots. Similarly, genetic considerations make it advisable to conduct many animal experiments on members of the same litter, but litters are generally of rather small size, so a number of litters must be used. Here the individual animal, like the plot in the agricultural example, is the experimental unit; while the litters, like the groups of similar plots, are called "blocks" of experimental units. In this terminology we may say that many experiments are block experiments, or even tautologically (if we allow that there may be only one block) that *all* experiments are block experiments. The effects of the variations between blocks need to be investigated only because we wish to eliminate them. In fact, blocks are a "nuisance" factor in the experiment.

We proceed to a formal investigation of block experiments, based on that of Tocher (1952).

38.15 Suppose that an experiment is carried out with the experimental units in b blocks, each containing k (>1) units, so that there are $bk = n$ observations in all. The experiment is to investigate a cross-classification or hierarchical classification (or a mixture of these) of certain factors of interest. Suppose that there are t distinct cells in the classification; for a two-way classification with r rows and c columns, e.g., we have $t = rc$. We shall call these the t "treatments." The problem is how to allocate the t treatments to the k units in each block.

We shall assume that no treatment is to be allocated to more than one unit in each block. (*) This is a reasonable assumption, since there seems little point in duplicating a treatment within a block, rather than using it in a further block, if more observations are required. This assumption implies that $t \geq k$.

38.16 The experiment may be completely described by defining a *treatment matrix* of order $(k \times t)$ for each block. For the j th block, t_j has its (l, i) th element equal to 1 or 0 according as the i th treatment is or is not allocated to the l th unit in that block. Evidently, there will be one non-zero element in each row (since one treatment is allocated to each unit) and no more than one in each column (because of the last assumption in 38.15).

If we require only to describe the allocation of treatments to blocks, without reference to their allocation to units within blocks, we can condense the information from the

(*) This assumption, which is satisfied by all well-known experiment designs, may be relaxed (cf. Tocher (1952)).

b treatment matrices \mathbf{t}_j into the *incidence matrix* of the experiment, \mathbf{n} , of order $(t \times b)$, which has its (i, j) th element n_{ij} equal to 1 or 0 according to whether or not the i th treatment occurs in the j th block.

38.17 If \mathbf{n}_j is the j th column of \mathbf{n} , and $\mathbf{1}_p$ is a $(p \times 1)$ vector of units, we have for this is simply summing each column of \mathbf{t}_j . Also, since all entries in \mathbf{t}_j are 0 or 1, we have

$$\mathbf{t}'_j \mathbf{1}_b = \mathbf{n}_j, \quad (38.1)$$

$$\mathbf{t}'_j \mathbf{t}_j = \text{diag}(\mathbf{n}_j) \quad (38.2)$$

where $\text{diag}(\mathbf{z})$ means a diagonal matrix with the vector \mathbf{z} as diagonal. If the i th treatment occurs r_i (>0) times in the experiment as a whole, and \mathbf{r} is the $(t \times 1)$ vector of the r_i , summing each row of \mathbf{n} gives

$$\mathbf{n} \mathbf{1}_b = \mathbf{r} \quad (38.3)$$

and summing each column gives

$$\mathbf{n}' \mathbf{1}_t = k \mathbf{1}_b, \quad (38.4)$$

since there are k units in each block. Further, (38.2-3) give

$$\sum_j \mathbf{t}'_j \mathbf{t}_j = \sum_j \text{diag}(\mathbf{n}_j) = \text{diag}(\sum_j \mathbf{n}_j) = \text{diag}(\mathbf{n} \mathbf{1}_b) = \text{diag}(\mathbf{r}). \quad (38.5)$$

38.18 In accordance with the spirit of our earlier discussion, the allocation of treatments to units will be randomized independently within each block, but we shall not for the present consider the effect of this within-blocks randomization upon the inferences drawn from the experiment—we return to this in 38.41. Here, we regard the randomization as a general precautionary measure against bias, and we conduct our analysis in terms of the linear model (Model I) familiar from earlier chapters.

Linear model for block experiments

38.19 An obvious linear model for a block experiment is

$$E(y_{ij}) = \tau_i + \beta_j, \quad (38.6)$$

where the τ_i are treatment effects and the β_j are block effects. Here, we are assuming that treatment and block effects are additive, with no interactions.

We saw at (19.19) that the only linear functions of parameters which can be unbiasedly estimated by linear functions of the observations are linear combinations of their expectations. Thus, in (38.6), only linear combinations of the $(\tau_i + \beta_j)$ can be so estimated, as is obvious from the fact that any constant added to all the τ_i and subtracted from all the β_j would leave (38.6) unaffected. We resolve this lack of identifiability of the τ_i , as we did in the singular model in 19.13-15, Vol. 2, and Example 19.9, by introducing a linear constraint upon the parameters. Since the block effects β_j are nuisance parameters, it is natural to impose the constraint upon them alone, in the form

$$\sum_{j=1}^b \beta_j = 0.$$

In effect, we add to the n y_{ij} a dummy random variable (zero), whose "expectation"

is the sum of the β_j , and this enables us to disentangle the unwanted block effects from the treatment effects in which we are interested.

38.20 Because of the constraint, we now require only $(b-1)$ parameters for a non-singular representation of the block effects. Call these $\alpha_1, \alpha_2, \dots, \alpha_{b-1}$, and array them as a vector α . We may obtain α from the vector β of the β_j by defining an orthogonal $(b \times b)$ matrix U_0 whose first row is $b^{-1/2} \mathbf{1}_b'$ and the remaining $(b-1)$ rows arbitrary, say \mathbf{u} . Then

$$U_0 \beta = \begin{pmatrix} 0 \\ \mathbf{u}\beta \end{pmatrix}, \quad (38.7)$$

since the constraint on the β_j is

$$\mathbf{1}_b' \beta = 0. \quad (38.8)$$

Because U_0 is orthogonal, $U_0^{-1} = U_0'$, so (38.7) gives

$$\beta = U_0' \begin{pmatrix} 0 \\ \mathbf{u}\beta \end{pmatrix} = \mathbf{u}' \mathbf{u} \beta. \quad (38.9)$$

Thus if we define

(38.7) and (38.9) become

$$\left. \begin{aligned} \alpha &= \mathbf{u}\beta, \\ U_0 \beta &= \begin{pmatrix} 0 \\ \alpha \end{pmatrix}, \\ \beta &= \mathbf{u}' \alpha. \end{aligned} \right\} \quad (38.10)$$

38.21 We now proceed to the formal LS solution for block experiments. We write \mathbf{y}_j for the vector of the k observations in the j th block, \mathbf{y} for the vector containing all the \mathbf{y}_j , and we partition \mathbf{u} into its column vectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_b$, each of order $(b-1)$. Then (38.6) may be written

$$\mathbf{y} = \mathbf{X}\theta + \epsilon \quad (38.11)$$

where \mathbf{X} and θ are conformably partitioned, with

$$\mathbf{X}_{bk \times (t+b-1)} = \begin{pmatrix} \mathbf{t}_1 & | & \mathbf{1}_k \mathbf{u}_1' \\ \mathbf{t}_2 & | & \mathbf{1}_k \mathbf{u}_2' \\ \vdots & | & \vdots \\ \mathbf{t}_b & | & \mathbf{1}_k \mathbf{u}_b' \end{pmatrix} \quad (38.12)$$

and

$$\theta_{(t+b-1) \times 1} = \begin{pmatrix} \tau \\ \alpha \end{pmatrix}. \quad (38.13)$$

The errors in (38.12) are as usual assumed to have mean zero, variance σ^2 , and to be uncorrelated.

38.22 From (38.13), we have

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \sum_j \mathbf{t}_j' \mathbf{t}_j & | & \sum_j \mathbf{t}_j' \mathbf{1}_k \mathbf{u}_j' \\ \hline \sum_j \mathbf{u}_j \mathbf{1}_k' \mathbf{t}_j & | & \sum_j \mathbf{u}_j \mathbf{1}_k' \mathbf{1}_k \mathbf{u}_j' \end{pmatrix} = \begin{pmatrix} \text{diag}(\mathbf{r}) & | & \mathbf{n}\mathbf{u}' \\ \hline \mathbf{u}\mathbf{n}' & | & k\mathbf{I}_{b-1} \end{pmatrix}, \quad (38.14)$$

on using (38.5), (38.1) and the relations $\mathbf{1}'_k \mathbf{1}_k = k$, $\mathbf{u}\mathbf{u}' = \mathbf{I}_{b-1}$. We may now invert (38.15). Writing

$$\Omega = \{\text{diag}(\mathbf{r}) - \mathbf{n}\mathbf{u}'\mathbf{u}\mathbf{n}'/k\}^{-1}, \quad (38.16)$$

we find

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \Omega & -\Omega\mathbf{n}\mathbf{u}'/k \\ -\mathbf{u}\mathbf{n}'\Omega/k & \{\mathbf{I}_{b-1} + \mathbf{u}\mathbf{n}'\Omega\mathbf{n}\mathbf{u}'/k\}/k \end{pmatrix} \quad (38.17)$$

as may be verified by multiplication of (38.15) and (38.17). Also

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum_j \mathbf{t}'_j \mathbf{y}_j \\ \sum_j \mathbf{u}_j \mathbf{1}'_k \mathbf{y}_j \end{pmatrix} = \begin{pmatrix} \mathbf{T} \\ \mathbf{u}\mathbf{B} \end{pmatrix}, \quad (38.18)$$

where $\mathbf{T} = \begin{pmatrix} T_1 \\ \vdots \\ T_t \end{pmatrix}$ and T_i is the total of y for all units receiving the i th treatment,

while $\mathbf{B} = \begin{pmatrix} B_1 \\ \vdots \\ B_b \end{pmatrix}$, where $B_j = \mathbf{1}'_k \mathbf{y}_j$, is the total of y for all units in the j th block.

From (38.17-18), the LS estimators are

$$\hat{\theta} = \begin{pmatrix} \hat{\tau} \\ \hat{\alpha} \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

where

$$\hat{\tau} = \Omega(\mathbf{T} - \mathbf{n}\mathbf{u}'\mathbf{u}\mathbf{B}/k) \quad (38.19)$$

and

$$\hat{\alpha} = \mathbf{u}(\mathbf{B} - \mathbf{n}'\hat{\tau})/k. \quad (38.20)$$

From (38.20) and (38.11), we then obtain for the original block parameters β ,

$$\hat{\beta} = \mathbf{u}'\hat{\alpha} = \mathbf{u}'\mathbf{u}(\mathbf{B} - \mathbf{n}'\hat{\tau})/k. \quad (38.21)$$

38.23 We now simplify the estimators (38.19) and (38.21) for computational purposes. First, since \mathbf{U}_0 in 38.20 is orthogonal, we have

$$\mathbf{I}_b = \mathbf{U}'_0 \mathbf{U}_0 = (b^{-\frac{1}{2}} \mathbf{1}'_b)' (b^{-\frac{1}{2}} \mathbf{1}_b) + \mathbf{u}'\mathbf{u}$$

so

$$\mathbf{u}'\mathbf{u} = \mathbf{I}_b - \mathbf{1}_b \mathbf{1}'_b / b. \quad (38.22)$$

Substitution of (38.22) into (38.16) gives, remembering (38.3),

$$\Omega = \{\text{diag}(\mathbf{r}) - \mathbf{n}\mathbf{n}'/k + \mathbf{r}\mathbf{r}'/(bk)\}^{-1}, \quad (38.23)$$

and (38.19) similarly becomes

$$\hat{\tau} = \Omega\{\mathbf{T} - \mathbf{n}\mathbf{B}/k + \mathbf{r}\mathbf{G}/(bk)\} \quad (38.24)$$

where we have written

$$\mathbf{G} = \mathbf{1}'_b \mathbf{B} = \mathbf{1}'_t \mathbf{T} \quad (38.25)$$

for the grand total of all the observations y . (38.24) may be further simplified, for (38.23) gives

$$\Omega^{-1} \mathbf{1}_t = \mathbf{r} - \mathbf{n}(\mathbf{n}' \mathbf{1}_t)/k + \mathbf{r}(\mathbf{r}' \mathbf{1}_t)/(bk). \quad (38.26)$$

Using (38.3-4) and their consequence

$$\mathbf{r}'\mathbf{1}_t = bk, \quad (38.27)$$

the last two terms in (38.26) cancel, and $\Omega^{-1}\mathbf{1}_t = \mathbf{r}$. Thus

$$\Omega\mathbf{r} = \mathbf{1}_t, \quad (38.28)$$

and (38.24) becomes

$$\hat{\tau} = \Omega(\mathbf{T} - \mathbf{nB}/k) + \mathbf{1}_t G/(bk). \quad (38.29)$$

Now we substitute (38.22) into (38.21) to obtain

$$\hat{\beta} = (\mathbf{I}_b - \mathbf{1}_b \mathbf{1}_b'/b)(\mathbf{B} - \mathbf{n}'\hat{\tau})/k. \quad (38.30)$$

This can again be simplified, for, using (38.25) and (38.3),

$$\begin{aligned} \mathbf{1}_b'(\mathbf{B} - \mathbf{n}'\hat{\tau}) &= G - \mathbf{r}'\hat{\tau}, \\ &= G - \mathbf{r}'\Omega(\mathbf{T} - \mathbf{nB}/k) - \mathbf{r}'\mathbf{1}_t G/(bk) \end{aligned} \quad (38.31)$$

on substituting (38.29). Now Ω is symmetric, so (38.28) implies $\mathbf{r}'\Omega = \mathbf{1}_t'$. If we now use (38.25), (38.4) and (38.27) in (38.31), we find that it reduces to zero. Thus (38.30) becomes simply

$$\hat{\beta} = (\mathbf{B} - \mathbf{n}'\hat{\tau})/k. \quad (38.32)$$

Apart from the calculation of Ω at (38.23), the treatment parameters estimator at (38.29) and the block parameters estimator at (38.32) require only the vectors of treatment and block totals, \mathbf{T} and \mathbf{B} , and the grand total G obtained from either of these by (38.25).

AV for block experiments

38.24 In order to construct the AV for the block experiment, we now require the Residual SS. This is

$$S_0 = (\mathbf{y} - \mathbf{X}\hat{\theta})'(\mathbf{y} - \mathbf{X}\hat{\theta}) \equiv \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\theta},$$

and using (38.18) and (38.11) this is

$$\begin{aligned} S_0 &= \mathbf{y}'\mathbf{y} - \begin{pmatrix} \mathbf{T} \\ \mathbf{uB} \end{pmatrix}' \begin{pmatrix} \hat{\tau} \\ \hat{\alpha} \end{pmatrix} = \mathbf{y}'\mathbf{y} - \mathbf{T}'\hat{\tau} - \mathbf{B}'\mathbf{u}'\hat{\alpha} \\ &= \mathbf{y}'\mathbf{y} - \mathbf{T}'\hat{\tau} - \mathbf{B}'\hat{\beta}. \end{aligned} \quad (38.33)$$

In general, the AV is non-orthogonal, since the off-diagonal matrices in (38.17) are non-null. We must therefore, as in Example 35.4, 35.38 and 35.43, find the Residual SS, say S_1 , when there are no treatment differences, only block parameters and a single treatment parameter being estimated. The difference $S_1 - S_0$ will then be the SS attributable to treatment differences.

38.25 We thus require to modify $\hat{\tau}$ so that it is of form $\hat{\tau}\mathbf{1}_t$. In (38.24), this gives

$$\hat{\tau}\mathbf{1}_t = \Omega\{\mathbf{T} - \mathbf{nB}/k + \mathbf{r}G/(bk)\}.$$

We substitute $\Omega\mathbf{r}$ for $\mathbf{1}_t$ from (38.28). Premultiplying by Ω^{-1} , transposing and postmultiplying by $\mathbf{1}_t$, we have

$$\hat{\tau}\mathbf{r}'\mathbf{1}_t = \{\mathbf{T} - \mathbf{nB}/k + \mathbf{r}G/(bk)\}'\mathbf{1}_t.$$

The first two terms on the right cancel, because (38.25) and (38.4) give

$$\mathbf{1}_t'(\mathbf{T} - \mathbf{nB}/k) = 0. \quad (38.34)$$

We thus find, using the remaining term on the right,

$$\hat{\tau} = G/(bk)$$

or

$$\hat{\tau}1_i = 1_i G/(bk), \quad (38.35)$$

a result intuitively obvious, since in the absence of treatment differences, the general mean $G/(bk)$ will be the estimator for all treatments. We now find from (38.32), using (38.35) and (38.4), that in this case

$$(\hat{\beta})\hat{\tau} = \hat{\tau}1_i = (B - 1_b G/b)/k. \quad (38.36)$$

Substituting (38.35-6) into (38.33), we find using (38.25) that

$$S_1 = y'y - B'B/k, \quad (38.37)$$

so that (as the reader is asked to verify in Exercise 38.1)

$$\begin{aligned} S_1 - S_0 &= T'\hat{\tau} + B'\hat{\beta} - B'B/k \\ &= (T - nB/k)' \Omega (T - nB/k) \end{aligned} \quad (38.38)$$

is the SS for treatment differences, while from (38.37) the combined SS for blocks and the general mean is $B'B/k$.

38.26 We may now display all these results in the AV table:

AV table for the general block experiment

Source of variation	SS	D.fr.
Treatment differences (allowing for block effects)	$T'\hat{\tau} + B'\hat{\beta} - B'B/k = (T - nB/k)' \Omega (T - nB/k)$	$t - 1$
Block effects (ignoring treatment differences)	$B'B/k - G^2/(bk)$	$b - 1$
Residual	$y'y - T'\hat{\tau} - B'\hat{\beta}$	$bk - b - t + 1$
General mean	$G^2/(bk)$	1
TOTAL	$y'y$	bk

(38.39)

The d.fr. for the Residual are obtained as a difference. (38.39) makes it clear that our analysis has simply separated off, from the d.fr. remaining after the $(b-1)$ linearly independent block parameters α and the general mean are allowed for, $(t-1)$ d.fr. for treatment differences.

The design of block experiments

38.27 The crucial computation in the preceding analysis is that of Ω at (38.23), requiring a matrix inversion. Ω depends upon the incidence matrix n and the vector r obtained from it by (38.3). Since $\sigma^2(X'X)^{-1}$ is, by (19.16), the dispersion matrix of $\hat{\theta}$, (38.17) shows that

$$V(\tau) = \sigma^2 \Omega. \quad (38.40)$$

We now seek to *design* the experiment (i.e. choose \mathbf{n}) so that the dispersion matrix of the treatment parameter estimators has some desired form. In determining this form we shall impose intuitively acceptable conditions, which will lead us to the most commonly used designs. Kiefer (1958, 1959) discusses various concepts of optimality in experiment design, and shows that designs which are symmetrical between all treatments (such as those to which we shall be led by imposing symmetrical conditions upon all treatment parameter estimators) are optimum in most of the senses.

In particular, these symmetrical designs minimize the generalized variance(*) (the determinant of (38.40)) and have optimum local power properties in testing the equality of all treatment parameters. Remarkably enough, these optimum properties are not retained if the design itself may be chosen by a random procedure, although the generalized variance is still minimized as block size $k \rightarrow \infty$.

For some theory and discussion of such *random allocation* (including *random balance*) designs, cf. Dempster (1960-1), Satterthwaite (1959), Budne (1959), the discussion of the last two papers by Youden *et al.* (1959), and Anscombe (1959).

38.28 If we choose \mathbf{n} to make $\mathbf{\Omega}$ diagonal, the treatment parameter estimators will be uncorrelated (orthogonal) and in addition the required matrix inversion will be trivial. $\mathbf{\Omega}$ will be diagonal if and only if its inverse is, and since the first term in the braces in (38.23) is already diagonal, we require only that $\mathbf{M} = \mathbf{nn}' - \mathbf{rr}'/b$ should be diagonal. Now, using (38.4) and (38.27),

$$\mathbf{1}_t' \mathbf{M} \mathbf{1}_t = \mathbf{1}_t' (\mathbf{nn}' - \mathbf{rr}'/b) \mathbf{1}_t = k \mathbf{1}_b' k \mathbf{1}_b - (bk)^2/b = 0. \quad (38.41)$$

Thus the sum of all the elements of \mathbf{M} is always zero. If \mathbf{M} is to be a diagonal matrix, its off-diagonal elements must all be zero, and therefore the sum of its diagonal elements, say M_{ii} , must be zero, i.e.

$$0 = \sum_{i=1}^t M_{ii} = \sum_{i=1}^t \left(\sum_{j=1}^b n_{ij}^2 - r_i^2/b \right) = \sum_{i=1}^t \sum_{j=1}^b (n_{ij} - r_i/b)^2 \quad (38.42)$$

on using (38.3). Thus we must have

$$n_{ij} - r_i/b = 0, \quad \text{all } i, j, \quad (38.43)$$

or in matrix terms

$$\mathbf{n} = \mathbf{r} \mathbf{1}_b' / b. \quad (38.44)$$

The condition for $\mathbf{\Omega}$ to be diagonal is thus that every block contains the same set of treatments, and every treatment is applied to a total number of units which is a multiple of b . Since each treatment occurs 0 or 1 times in any block, and no treatment occurs 0 times in the experiment as a whole, this implies that each treatment occurs once in each block. Thus $k = t$ and

$$\mathbf{r} = b \mathbf{1}_t, \quad (38.45)$$

so (38.44) becomes simply

$$\mathbf{n} = \mathbf{1}_t \mathbf{1}_b'. \quad (38.46)$$

(*) Within any design, we know from 19.8, Vol. 2, that the LS estimators minimize the generalized variance among linear estimators. The result above refers to the choice *between* designs.

This is the incidence matrix of a *randomized blocks* design.
Exercises 38.2-3 give related results.

Randomized blocks designs

38.29 We have therefore arrived, by a lengthy route, at the randomized blocks designs which were briefly mentioned in 36.39, when we first discussed the randomized allocation of units to treatments. The structure of a randomized block experiment is extremely simple: t treatments are randomly allocated to the $k = t$ units in each of b blocks. Because of the diagonality of Ω from which the designs were deduced in 38.28, the LS estimators of the parameters and the general AV table for block experiments at (38.39) simplify greatly. Using (38.45-6) in (38.23) we find

$$\Omega = \mathbf{I}_t/b, \quad (38.47)$$

while use of (38.46) and (38.25) in (38.29) and (38.32) gives

$$\hat{\tau} = \mathbf{T}/b, \quad (38.48)$$

$$\hat{\beta} = \mathbf{B}/t - \mathbf{1}_b G/(bt). \quad (38.49)$$

Similarly, the SS for treatment differences (38.38) becomes

$$S_1 - S_0 = \mathbf{T}'\mathbf{T}/b - G^2/(bt) \quad (38.50)$$

and the Residual SS (38.33) becomes

$$S_0 = \mathbf{y}'\mathbf{y} - \mathbf{T}'\mathbf{T}/b - \mathbf{B}'\mathbf{B}/t + G^2/(bt). \quad (38.51)$$

The reader is asked to verify these formulae in Exercise 38.4.

38.30 We now display the simplified AV table:

AV table for a randomized blocks experiment

Source of variation	SS	D.fr.
Treatment differences	$\mathbf{T}'\mathbf{T}/b - G^2/(bt)$	$t - 1$
Block differences	$\mathbf{B}'\mathbf{B}/t - G^2/(bt)$	$b - 1$
Residual	$\mathbf{y}'\mathbf{y} - \mathbf{T}'\mathbf{T}/b - \mathbf{B}'\mathbf{B}/t + G^2/(bt)$	$(t - 1)(b - 1)$
General mean	$G^2/(bt)$	1
TOTAL	$\mathbf{y}'\mathbf{y}$	bt

(38.52)

Comparison with (38.39) reveals the extent of the simplification.

The symmetry of the table (38.52) as between treatments (the symbols \mathbf{T} , t) and blocks (the symbols \mathbf{B} , b) makes it clear that blocks are in fact being treated as a (nuisance) factor in the analysis, for (38.52) is formally identical with the AV table for the two-way cross-classification with one observation per cell (cf. Example 35.3 and Exercise 35.1). As always in that analysis, there is no Interactions SS, though we can separate

off a single d.fr. from the Residual SS to test Interactions as in Example 35.3. Thus we can check the assumption in the model that treatments and blocks do not interact.

Example 38.2

In 38.10 we saw that neglect of the time of day at which the tests were taken, in the experiment of Example 38.1, renders that experiment vulnerable to criticism. If now we treat time of day as a nuisance factor, and arrange the experiment in randomized blocks (each block consisting of a single time of day), this criticism is met.

Economizing degrees of freedom: two nuisance factors

38.31 It will be seen from (38.39) that $(b-1)$ d.fr. are absorbed by the b blocks used. If it is necessary, owing to great variability between units, even within blocks, to keep the number of units per block very small, the number of blocks must be correspondingly increased to obtain a given number of observations. Thus a large number of degrees of freedom will be absorbed by the blocks, and it may be necessary to seek an alternative experiment design to avoid this—a situation with b much greater than t clearly leaves much to be desired if bt is small.

To economize the number of d.fr. lost in eliminating blocks, we must arrange that the blocks do not each require a parameter as in 38.19. A simple way of achieving this is to arrange the blocks themselves in a two-way classification, say with n rows and m columns; there are k units in each block as before. We now assume that the $nm = b$ block parameters β_{ij} satisfy

$$\beta_{ij} = \rho_i + \gamma_j, \quad i = 1, 2, \dots, n; j = 1, 2, \dots, m. \quad (38.53)$$

Thus we are, in effect, using the blocks to eliminate the effects of two nuisance factors, corresponding to the row- and column-classifications of the blocks. (38.53) is evidently a restrictive model—it amounts to assuming that the two nuisance factors have main effects, but no interactions. If the latter assumption is false, the analysis is invalid.

As before, we impose the constraints

$$\sum_{i=1}^n \rho_i = \sum_{j=1}^m \gamma_j = 0 \quad (38.54)$$

to ensure that the treatment effects can be identified. There will thus be only $(n-1)+(m-1)$ d.fr. absorbed by the blocks.

We may now sketch the LS analysis, which is quite analogous to that of 38.15–26.

38.32 There are now nm treatment matrices t_{ij} as in 38.16. We let \mathbf{n} be the incidence matrix of order $(t \times n)$ relating to the rows, so that n_{il} is 1 or 0 according as the i th treatment occurs in the l th row (irrespective of column); similarly, \mathbf{m} is the incidence matrix for columns, of order $(t \times m)$. We then have, as at (38.3),

$$\mathbf{n}\mathbf{l}_n = \mathbf{m}\mathbf{l}_m = \mathbf{r}. \quad (38.55)$$

To implement (38.54), (38.8) is replaced by

$$\mathbf{l}'_n \boldsymbol{\rho} = \mathbf{l}'_m \boldsymbol{\gamma} = 0, \quad (38.56)$$

and instead of $\boldsymbol{\alpha}$ defined at (38.10), we have

$$\left. \begin{aligned} \boldsymbol{\sigma} &= \mathbf{u}\boldsymbol{\rho} \\ \boldsymbol{\delta} &= \mathbf{v}\boldsymbol{\gamma} \end{aligned} \right\} \quad (38.57)$$

where σ is of order $(n-1)$, δ of order $(m-1)$ and \mathbf{u} , \mathbf{v} are analogues of \mathbf{u} in (38.12) now holds with \mathbf{y} a vector containing nm vectors \mathbf{y}_{ij} ,

$$\mathbf{X}_{nmk \times (t+n+m-2)} = \begin{pmatrix} \mathbf{t}_{11} & \mathbf{1}_k \mathbf{u}'_1 & \mathbf{1}_k \mathbf{v}'_1 \\ \vdots & \vdots & \vdots \\ \mathbf{t}_{lj} & \mathbf{1}_k \mathbf{u}'_l & \mathbf{1}_k \mathbf{v}'_j \\ \vdots & \vdots & \vdots \\ \mathbf{t}_{nm} & \mathbf{1}_k \mathbf{u}'_n & \mathbf{1}_k \mathbf{v}'_m \end{pmatrix} \quad (38.58)$$

and

$$\boldsymbol{\theta}_{(t+n+m-2) \times 1} = \begin{pmatrix} \tau \\ \sigma \\ \delta \end{pmatrix}. \quad (38.59)$$

We now find that (cf. (38.15))

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \text{diag}(\mathbf{r}) & \mathbf{nu}' & \mathbf{mv}' \\ \mathbf{un}' & mk\mathbf{I}_{n-1} & \mathbf{0} \\ \mathbf{vm}' & \mathbf{0} & nk\mathbf{I}_{m-1} \end{pmatrix}, \quad (38.60)$$

and if we write, analogously to (38.23),

$$\boldsymbol{\Omega} = \{\text{diag}(\mathbf{r}) - \mathbf{nn}'/(mk) - \mathbf{mm}'/(nk) + 2\mathbf{rr}'/(nmk)\}^{-1}, \quad (38.61)$$

the inverse of (38.60) is

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \boldsymbol{\Omega} & -\boldsymbol{\Omega}\mathbf{nu}'/(mk) & -\boldsymbol{\Omega}\mathbf{mv}'/(nk) \\ -\mathbf{un}'\boldsymbol{\Omega}/(mk) & \{\mathbf{I}_{n-1} + \mathbf{un}'\boldsymbol{\Omega}\mathbf{nu}'/(mk)\}/(mk) & \mathbf{un}'\boldsymbol{\Omega}\mathbf{mv}'/(nmk^2) \\ -\mathbf{vm}'\boldsymbol{\Omega}/(nk) & \mathbf{vm}'\boldsymbol{\Omega}\mathbf{nu}'/(nmk^2) & \{\mathbf{I}_{m-1} + \mathbf{vm}'\boldsymbol{\Omega}\mathbf{mv}'/(nk)\}/(nk) \end{pmatrix}, \quad (38.62)$$

so that (38.40) remains true. Just as at (38.18), we may write

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} \mathbf{T} \\ \mathbf{uR} \\ \mathbf{vC} \end{pmatrix} \quad (38.63)$$

where \mathbf{T} , \mathbf{R} and \mathbf{C} are respectively the vectors of treatment, row and column totals. G is the grand total as before.

38.33 Multiplication of (38.62) by (38.63) gives the LS estimators. Instead of (38.19), we now find

$$\hat{\tau} = \boldsymbol{\Omega}\{\mathbf{T} - \mathbf{nu}'\mathbf{uR}/(mk) - \mathbf{mv}'\mathbf{vC}/(nk)\} \quad (38.64)$$

which simplifies as at (38.24) to

$$\hat{\tau} = \boldsymbol{\Omega}\{\mathbf{T} - \mathbf{nR}/(mk) - \mathbf{mC}/(nk) + 2\mathbf{rG}/(mnk)\}. \quad (38.65)$$

As at (38.32), we find

$$\begin{cases} \hat{\rho} = (\mathbf{R} - \mathbf{n}'\hat{\tau})/(mk), \\ \hat{\gamma} = (\mathbf{C} - \mathbf{m}'\hat{\tau})/(nk). \end{cases} \quad (38.66)$$

The Residual SS is therefore, as at (38.33),

$$S_0 = \mathbf{y}'\mathbf{y} - \mathbf{T}'\hat{\boldsymbol{\tau}} - \mathbf{R}'\hat{\boldsymbol{\rho}} - \mathbf{C}'\hat{\boldsymbol{\gamma}}$$

(38.67)

(38.68)

(38.69)

and as in 38.25, if we find that the Residual SS is

$$S_1 = \mathbf{y}'\mathbf{y} - \mathbf{R}'\mathbf{R}/(mk) - \mathbf{C}'\mathbf{C}/(nk) + G^2/(nmk),$$

(38.70)

so that the SS for treatment differences is, from (38.67) and (38.69),

$$S_1 - S_0 = \mathbf{P}'\boldsymbol{\Omega}\mathbf{P} - G^2/(nmk),$$

where we now write \mathbf{P} for the matrix in braces on the right of (38.65). (38.70) is the analogue of (38.38). The AV table may now be constructed just as at (38.39), with $(n+m-2)$ d.fr. in all for the block effects. We leave this, and the verification of the formulae above, to the reader as Exercise 38.5.

Latin squares

38.34 The elimination of two nuisance factors by the method 38.31-3 is most commonly effected with $n = m$ and $k = 1$. Each block then consists of a single unit, and the units are in a square array. In 38.31-3, we have said nothing about how the t treatments are to be allocated to the blocks. We now assume that $t = n = m$, so that we have an array of t^2 units to which t treatments are to be randomly allocated. If we impose the condition that each treatment be allocated just once to each row and just once to each column of the array, we obtain the arrangement known as a *Latin square*. It is exemplified by (38.71), with $t = 4$ and the treatments labelled as A, B, C, D :

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

(38.71)

Euler studied Latin squares extensively from a purely mathematical viewpoint in the eighteenth century. The fact that they have come to be useful in the design of experiments in the present century is a notable example of the possible ultimate practical value of apparently useless theorizing.

38.35 If we look only at the rows of (38.71), we have a randomized blocks design with $b = t$; and similarly if we look only at the columns. We may thus use the AV table (38.52) with $b = t$ and *either* set of block differences in the second row of the table. This leads us to expect that we should be able to separate off *both* sets of block differences to obtain an AV table which in outline is:

Source of variation	D.fr.
Treatment differences	$t - 1$
Rows	$t - 1$
Columns	$t - 1$
Residual	$t - 1$
General mean	$(t - 1)(t - 2)$
TOTAL	1
	t^2

(38.72)

In fact, this is immediately deducible from the results of 38.33, by putting $n = m = t$ and $k = 1$. We leave this to the reader as Exercise 38.6.
If the two nuisance factors interact, contrary to the assumption of their additivity made at (38.53), the analysis may be seriously in error (cf. Scheffé (1959)).

38.36 For a randomized blocks design, there is no difficulty in choosing one at random—we require only random permutations of numbers from 1 to t (cf. 9.14 and Example 9.7) to choose with equal probabilities from among the $(t!)^b$ possible allocations of experimental units to treatments.

For Latin squares, however, the choice of a design at random is less straightforward, since it is not at all obvious how many squares of a given order exist, though it is evident from consideration of the cyclic permutations of the elements of the first row that some Latin squares of any order do exist.

The numbers of possible Latin squares of order t is very large for high values of t . There are, for example, 576 squares of order 4; 161,280 squares of order 5; and 812,851,200 of order 6. Up to order 7, they have been counted. Although many examples of squares of higher orders are known, the problem of enumeration for $t \geq 8$ awaits solution. Details and examples will be found in Fisher and Yates' *Statistical Tables*.

By interchanging rows and columns the square can always be brought to a form in which the top row and left-hand column are in the order ABC, etc. It is then said to be a "standard square." For instance, there are four standard squares of the fourth order:

A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	
B	A	D	C	B	C	D	A	B	D	A	C	B	A	D	C	
C	D	B	A	C	D	A	B	C	A	D	B	C	D	A	B	(38.73)
D	C	A	B	D	A	B	C	D	C	B	A	D	C	B	A	

From each of these, $144 (= 4!3!)$ squares may be derived by permuting all columns, and all rows except the first. (There is no point in permuting the first row, because the result would be a repetition of squares already obtained with an interchange of the letters $A \dots D$, not an essentially different layout.) The total number of squares, as stated above, is therefore $4 \times 144 = 576$. More generally, each standard square yields $t!(t-1)!$ squares of order t .

It is thus only necessary to specify the standard squares. To select a Latin square at random, we choose a standard form at random and then permute rows and columns at random, the randomizing process being most conveniently carried out by using tables of random permutations (cf. 9.14 and Example 9.7). For squares of order 8 or more, where the standard types have not been enumerated, we can only choose one of those which has, and hence select one at random from a restricted set of all possible squares.

Three or more nuisance factors: Graeco-Latin and orthogonal squares

38.37 There is no difficulty in generalizing the Latin square to provide for the elimination of three or more nuisance factors. We can do this by a process of

superposition of different Latin squares. If the Latin square

$$\begin{array}{cccc} A & B & C & D \\ C & D & A & B \\ D & C & B & A \\ B & A & D & C \end{array} \quad (38.74)$$

is superposed upon the square (38.71), with the Roman letters of (38.74) first changed to the corresponding Greek letters to avoid confusion, we obtain the arrangement

$$\begin{array}{cccc} A\alpha & B\beta & C\gamma & D\delta \\ B\gamma & A\delta & D\alpha & C\beta \\ C\delta & D\gamma & A\beta & B\alpha \\ D\beta & C\alpha & B\delta & A\gamma \end{array} \quad (38.75)$$

in which each combination of a Roman with a Greek letter appears just once. The squares (38.71) and (38.74) are said to be *orthogonal (Latin) squares* for this reason. Their superposition (38.75) is called a *Graeco-Latin square*.

There is evidently no Graeco-Latin square when $t = 2$. More remarkably, there is none when $t = 6$ even though there are 812,851,200 Latin squares for $t = 6$. Euler conjectured that no Graeco-Latin square exists when $t = 4r + 2$, and it has taken nearly two centuries to disprove his conjecture and show (Bose and Shrikande, 1959, 1960; Bose *et al.*, 1960) that a Graeco-Latin square exists except only when $t = 2$ or 6. Fisher and Yates' *Tables* give examples.

The Greek letters in (38.75) may be used to identify a third nuisance factor (rows and columns, as before, identifying the first two), while the Roman letters are the treatments as before. The design then eliminates the effects of three nuisance factors in exactly the same way as the Latin square eliminates the effects of two. The AV table is an obvious generalization of Exercise 38.6, which we leave to the reader there.

38.38 A further Latin square (using a third set of symbols, say numerals) may be superposed upon (38.75) so that each combination of any two sets of symbols occurs just once, and the three Latin squares are mutually orthogonal. This is true for the arrangement

$$\begin{array}{cccc} A\alpha 1 & B\beta 2 & C\gamma 3 & D\delta 4 \\ B\gamma 4 & A\delta 3 & D\alpha 2 & C\beta 1 \\ C\delta 2 & D\gamma 1 & A\beta 4 & B\alpha 3 \\ D\beta 3 & C\alpha 4 & B\delta 1 & A\gamma 2 \end{array} \quad (38.76)$$

which is called a *hyper-Graeco-Latin square*. If the Greek letters and numerals are used to identify the third and fourth nuisance factors, this design will eliminate the effects of four nuisance factors. The AV is again left to the reader in Exercise 38.6.

38.39 No further Latin square can be superposed upon (38.76) while maintaining orthogonality; no more than $(t-1)$ Latin squares of order t can achieve this—the simple proof is left to the reader as Exercise 38.7. A set of $(t-1)$ orthogonal squares of order t , like (38.76), is called a *complete set* of orthogonal Latin squares. Such

complete sets exist if t is prime or a power of a prime (cf. Mann (1949)), and hence for all $t \leq 9$ except 2 and 6, when we have seen in 38.37 that not even a pair of orthogonal squares exists. For $10 \leq t \leq 30$, Barra (1965) gives details of the known sets of orthogonal squares. The complete sets have been enumerated for $t \leq 7$. Fisher and Yates' *Tables* give examples for $t \leq 9$.

For details of the theory and construction of Latin squares and orthogonal sets of them, into which we shall not enter here, the reader should refer to Mann's (1949) account of the Galois field methods due to R. C. Bose, and to the review paper by Barra (1965), which contains a bibliography of the subject subsequent to the earlier review by Norton (1939).

38.40 The practical usefulness of Latin squares in experimental work is restricted by the condition that the number of treatments must be the same as the number of levels for each nuisance factor, and this restriction increases as we pass through the Graeco-Latin to the higher-order sets of orthogonal squares. In consequence, these latter arrangements are little used. However, Latin squares are frequently used in agricultural experimentation, where the rows and columns of the square array represent the physical rows and columns in which the experimental plots (units) are laid out. In this way, soil fertility gradients across the experimental area in these two directions will have no effect on the treatments. Of course, there may be fertility gradients in other directions, e.g. diagonally to the square array, which the Latin square arrangement does not eliminate. The experimenter will, however, choose the orientation of his rows and columns to eliminate known or likely fertility gradients.

It is clear that Latin square arrangements are of possible use whenever there are two geographical or temporal co-ordinates to be eliminated, and similarly that the higher-order arrangements may be called on if there are three or more such nuisance factors.

Example 38.3

In the experiment discussed in Examples 38.1–2, it might be suspected that the day of the week on which the tests are carried out also influences the result. The hypothesis here would be that there is some kind of cumulative fatigue through the working week, acting similarly to the "time of day" effect already discussed. We can eliminate the effect of both these nuisance factors by choosing as many times of day as there are days in the working week (say 5) and arranging the experiment in a 5×5 Latin square design. Notice that only 5 treatments (alcohol-doses) would be possible if we used a single square. There is, however, nothing to prevent our using as many 5×5 squares as are required to test all the proposed treatments, so long as we make the latter a multiple of 5.

If, in addition to time of day and day of the week, place of work were regarded as a third nuisance factor influencing the experiment, we could choose the participating drivers from five work-places and arrange the experiment as a Graeco-Latin square. But it is precisely the inflexibility of having to choose five work-places and five times of day, merely because the number of working days is fixed at five, which often makes these designs inconvenient.

Randomization and robustness in randomized blocks and Latin squares

38.41 In 38.18 we left aside the question of the effect upon inference of the random allocation of treatments to units within the blocks of a block experiment. The same question arises with respect to the random allocation of treatments to units in a Latin square (cf. 38.34). We must now examine this question in some detail.

The question is a double-sided one. In the first place, since we are (as a general precaution against biased inference) carrying out certain physical acts of randomization when we allocate treatments to experimental units, we may ask how this will affect the validity of the inferences drawn from a linear model with homoscedastic uncorrelated errors, which become independent when (as we must for hypothesis-testing) we assume their normality. Put this way, it is a question concerning the robustness of our inferences.

However, the question may be put more directly and ambitiously. The randomizations generate equiprobable sets of observations, and we have seen in Chapter 31 and subsequently that consideration of permutation distributions can permit us to replace normal distribution theory by more general (distribution-free) methods. These often lose little or no efficiency even when the normality assumption is valid, and may be much more efficient when it is not. We may ask whether randomization can here perform the same service of freeing us from dependence on the normality assumption.

38.42 A detailed account of randomization theory in randomized blocks and Latin squares is contained in the penultimate chapter of Scheffé (1959), which gives references to the literature. From the point of view of estimation, the most interesting results are those for the expected values of MS in the AV tables, quoted from Kempthorne (1952) and from Wilk and Kempthorne (1957) (cf. also D. R. Cox (1958b)).

For the randomized blocks design, the expected MS for treatments is less than that for the Residual by a term depending upon the interactions between blocks and treatments, as well as exceeding it by the usual term depending upon treatment effects. (It is noteworthy that the presence of interactive errors (cf. 36.41) between treatments and units *within* blocks does not affect the situation.) The Residual MS is thus inflated by blocks-treatments interactions. However, this difficulty disappears if (as is often appropriate) block effects are treated as random, rather than fixed, effects; the Residual MS may then properly be used to assess the magnitude of the MS for treatments.

For the Latin square, the situation is more complicated, for here interactive errors do have some effect upon the comparison of the MS for treatments with that for Residual. No simple result emerges, for essentially the reason discussed in 36.42.

38.43 We now consider the testing of treatment effects under randomization theory. For the case where there are no unit errors (cf. 36.41) we have already developed the theory in 37.39-41, where we were concerned with a permutation test for column-effects in a two-way classification with one observation per cell. We have seen in 38.30 that the randomized blocks situation is formally identical with such a classification. Thus the results of 37.39-41 hold for testing treatment effects in randomized blocks, the rows there being interpreted as blocks and the columns as treatments. We therefore

see that we may apply the usual AV test for treatments in (38.52), with d.f. adjusted by the method of 37.29 as indicated in 37.40. (A few sampling experiments by Collier and Baker (1966) indicate that the power of the usual F -test is also robust to non-normality.) Alternatively, we may use distribution-free tests based on ranks or normal scores as in 37.41 and Exercise 37.14, with negligible correction to the d.f. if either the number of treatments or the number of blocks is not too small.

If there are unit errors, Ogawa (1961, 1963) shows that the standard F -test may still be justified as an approximation if the variances of unit effects within blocks are nearly constant and the number of blocks is large enough.

Even in the absence of unit errors, the permutation test for treatment effects in Latin squares is less satisfactory than that for randomized blocks, just as we saw for estimation in 38.42. As before, the expected value of the usual AV test statistic is the same under randomization as under normal theory, but the variance is complicated, and in consequence the evidence for the approximation of the permutation distribution by normal theory is very limited (Welch, (1937); Scheffé (1959)).

38.44 The fact that the evidence for the validity of normal theory tests in randomized Latin squares is flimsy, together with the even greater paucity of such evidence for most other, more complicated, experiment designs, leads one to doubt the prevailing serene assumption that randomization theory will always approximate normal theory.

There is a question of principle involved here. Is randomization to be explicitly incorporated into the theory underlying our tests and estimation procedures? Since randomization lies at the root of the modern approach to statistical inference (cf. 38.7), it seems difficult not to answer this question in the affirmative, and consequently difficult to defend the relative neglect of this admittedly complicated branch of distribution theory.

The variances of treatment differences in block experiments

38.45 We now return to the problem of design in the general analysis of block experiments given in 38.19–26 above. Instead of requiring, as we did in 38.27, that the treatment parameter estimators should be orthogonal (leading to the randomized blocks design, as we found in 38.28–30), we now formulate the design problem in terms of the variances of the differences between these estimators.

Suppose that we wish the experiment to result in the variance of the difference between the i th and l th treatment parameter estimators being $2\sigma^2 d_{il}$, say. We write $\sigma^2 w_{il}$ for the (i, l) th element of the dispersion matrix of the treatment parameter estimators, equal to $\sigma^2 \Omega$ by (38.40) with Ω defined by (38.23). Then

$$w_{ii} - 2w_{il} + w_{ll} = 2d_{il} \quad (38.77)$$

so that

$$w_{il} = \frac{1}{2}(w_{ii} + w_{ll}) - d_{il}. \quad (38.78)$$

If we write \mathbf{w} for the vector with elements $\{\frac{1}{2}w_{ii}\}$ and \mathbf{D} for the matrix with elements $\{-d_{il}\}$, with $d_{ii} \equiv 0$ by (38.77), we may write (38.78) in the form

$$\Omega = \mathbf{w}\mathbf{1}' + \mathbf{1}\mathbf{w}' + \mathbf{D} = (\mathbf{w} \mid \mathbf{1}_l) \begin{pmatrix} \mathbf{1}'_l \\ -\mathbf{w}' \end{pmatrix} + \mathbf{D}. \quad (38.79)$$

Inspection of (38.79) makes it clear that it is identical with

$$\Omega = (\mathbf{w} - \frac{1}{2}c\mathbf{1}_t)\mathbf{1}_t' + \mathbf{1}_t(\mathbf{w} - \frac{1}{2}c\mathbf{1}_t)' + (\mathbf{D} + c\mathbf{1}_t\mathbf{1}_t') \quad (38.80)$$

whatever the scalar c may be. Since \mathbf{w} (a function of Ω) is itself at choice in the design of the block experiment, no generality is lost by continuing to use (38.79), with the proviso that \mathbf{D} therein may be replaced by any matrix which differs from \mathbf{D} by a scalar multiple of $\mathbf{1}_t\mathbf{1}_t'$. Such a matrix is said to be of class \mathbf{D} .

38.46 Suppose now that \mathbf{D} , a constant times the desired matrix of the variances of treatment differences, has all its off-diagonal elements equal, so that we desire all differences to be estimated with the same precision. (The leading diagonal of \mathbf{D} , of course, contains zeros.) This can, by (38.77), be achieved by choosing the dispersion matrix Ω of the treatment parameter estimators so that its diagonal elements w_{ii} are all equal, and its off-diagonal elements w_{it} are all equal. In this case, \mathbf{w} is itself a scalar multiple of $\mathbf{1}_t$ and we can choose c in (38.80) so that

$$\mathbf{w} - \frac{1}{2}c\mathbf{1}_t = \mathbf{0}.$$

(38.80) then becomes

$$\Omega = \mathbf{D}_c \quad (38.81)$$

where \mathbf{D}_c stands for a particular matrix of class \mathbf{D} . (38.81) and (38.23) give

$$\mathbf{D}_c^{-1} = \text{diag}(\mathbf{r}) - \mathbf{nn}'/k + \mathbf{rr}'/(bk)$$

or

$$\mathbf{nn}' = k\{\text{diag}(\mathbf{r}) - \mathbf{D}_c^{-1} + \mathbf{rr}'/(bk)\}. \quad (38.82)$$

(38.28) and (38.81) give

$$\mathbf{r} = \mathbf{D}_c^{-1}\mathbf{1}_t \quad (38.83)$$

and (38.83) used in (38.27) gives

$$bk = \mathbf{1}_t'\mathbf{D}_c^{-1}\mathbf{1}_t. \quad (38.84)$$

Substitution of (38.83-4) into (38.82) gives the alternative form

$$\mathbf{nn}' = k\left\{\text{diag}(\mathbf{D}_c^{-1}\mathbf{1}_t) - \mathbf{D}_c^{-1} + \frac{\mathbf{D}_c^{-1}\mathbf{1}_t\mathbf{1}_t'\mathbf{D}_c^{-1}}{\mathbf{1}_t'\mathbf{D}_c^{-1}\mathbf{1}_t}\right\}. \quad (38.85)$$

Although we have derived (38.85) under the special assumption that \mathbf{D} has all off-diagonal elements equal, it holds quite generally for any \mathbf{D} , as Tocher (1952) showed directly (cf. Exercise 38.8).

38.47 Still considering the special assumption of 38.46, we see that if the off-diagonal elements of \mathbf{D} are all equal to $-\frac{1}{a}$ (corresponding to variances of differences all equal to $2\sigma^2/a$), we have

$$\mathbf{D} = \frac{1}{a}(\mathbf{I}_t - \mathbf{1}_t\mathbf{1}_t') \quad (38.86)$$

and since

$$\begin{aligned} \mathbf{D}_c &= \mathbf{D} + c\mathbf{1}_t\mathbf{1}_t', \\ \mathbf{D}_c &= \frac{1}{a}\{\mathbf{I}_t + (ac - 1)\mathbf{1}_t\mathbf{1}_t'\}. \end{aligned} \quad (38.87)$$

The inverse of (38.87) may be verified to be

$$\mathbf{D}_c^{-1} = a \left\{ \mathbf{I}_t - \frac{(ac-1)}{1+t(ac-1)} \mathbf{1}_t \mathbf{1}_t' \right\}$$

which we simplify to

$$\mathbf{D}_c^{-1} = a \{ \mathbf{I}_t + A \mathbf{1}_t \mathbf{1}_t' \}, \quad (38.88)$$

where $A = (1-ac)/\{1+t(ac-1)\}$. We now find for (38.83)

$$\mathbf{r} = \mathbf{D}_c^{-1} \mathbf{1}_t = a(1+At) \mathbf{1}_t = r \mathbf{1}_t \quad (38.89)$$

where $r = a(1+At)$. We thus see that each treatment occurs in the experiment the same number, r , of times. Using (38.88-9), we find for (38.85)

$$\mathbf{nn}' = k \{ (r-a) \mathbf{I}_t + (a/t) \mathbf{1}_t \mathbf{1}_t' \}. \quad (38.90)$$

It will be seen that the arbitrary constant A has now disappeared; it is only relevant in determining r .

The design equation

38.48 An equation for \mathbf{nn}' is called a *design equation*. Because of the definition of \mathbf{n} in 38.16, we see that the (i, l) th element of \mathbf{nn}' counts the number of times (say λ_{il}) in the experiment as a whole that the i th and l th treatments occur together in the same block. In particular, when $i = l$, the diagonal elements of \mathbf{nn}' are simply the frequencies r_i with which the i th treatment occurs in the experiment.

Balanced incomplete blocks designs

38.49 We now interpret the particular design equation (38.90) in the light of 38.48. Equating its diagonal elements, we have

$$\{\mathbf{nn}'\}_{ii} = r = k \{ (r-a) + a/t \}$$

whence

$$r = ak(t-1)/\{(k-1)t\}. \quad (38.91)$$

The off-diagonal elements of (38.90) are

$$\{\mathbf{nn}'\}_{il} = ka/t = \lambda, \quad (38.92)$$

say. Thus, in the whole experiment, every pair of treatments occurs λ times together in the same block.

(38.91-2) give

$$r(k-1) = \lambda(t-1). \quad (38.93)$$

Since $t \geq k$, (38.93) implies $\lambda \leq r$, as is obvious from their definitions. Moreover, since each of t treatments occurs r times in the experiment with b blocks each containing k units, we have

$$rt = bk. \quad (38.94)$$

Using (38.92-3), (38.90) may be written

$$\mathbf{nn}' = (r-\lambda) \mathbf{I}_t + \lambda \mathbf{1}_t \mathbf{1}_t'. \quad (38.95)$$

If $r = \lambda$, (38.95) is satisfied by the randomized blocks design incidence matrix (38.46), for then $k = t$, $\lambda = b$ from (38.93-4); this is obvious from the definitions of r and λ . We henceforth exclude this case, and consider only $r > \lambda$, $t > k$.
A block experiment satisfying the design equation (38.95) and the conditions

(38.93-4) is called a *balanced incomplete blocks* (BIB) design. These designs were introduced by Yates (1936a). Their characterizing features are that each of the t treatments appears in r of the b blocks of k units, while each pair of the treatments appear together in λ of the blocks. We were led to BIB designs by the requirement in 38.46-7 that every treatment difference be estimated with the same variance $2\sigma^2/a$, where a is given by (38.91-2).

The conditions (38.93-4) show that any three of the five constants t, k, λ, r, b determine the other two; commonly, the first three of these constants are used.

38.50 Although (38.93-5) are necessary conditions (to be satisfied by integers t, k, λ, r, b) for a BIB design to exist, they are not in general sufficient conditions, since there may be no incidence matrix \mathbf{n} satisfying the design equation (38.95) for \mathbf{nn}' (cf. Exercise 38.11). Further necessary conditions may be given. For example, (38.95) implies that the $(t \times t)$ matrix \mathbf{nn}' is non-singular, so the $(t \times b)$ matrix \mathbf{n} must have rank t and

$$b \geq t \quad (38.96)$$

(which with (38.94) implies $r \geq k$), a result originally differently found by R. A. Fisher. In a comprehensive review of the subject, Guérin (1965) summarizes other, more stringent, necessary conditions. (38.93-5) are known to be sufficient, as well as necessary, for $k = 3$ or 4 , and also for $k = 5$ with $\lambda = 1, 4$ or 20 , results due to Hanani (1961).

For a given BIB design, λ can always be increased by any integral multiple m by simply repeating the whole design m times— r and b are then similarly increased. We shall always take $m = 1$ to make λ as small as possible for any given design. However, there may be more than one such “minimal” λ for given (t, k) , corresponding to different BIB designs—examples appear in the table in 38.54 below. Further, several BIB designs of essentially different structure (non-isomorphic) may exist for the same values of t, k and λ . This is intuitively obvious from the fact that the values of r and λ are first- and second-order conditions only upon the disposition of the treatments into the blocks—they do not in general restrict the frequencies of triples, quadruples, etc., of treatments.

38.51 If t and k are fixed, the third constant of the design being at choice, a BIB design can always be formed simply by taking every one of the $\binom{t}{k}$ selections of treatments as a block. We then have

$$b = \binom{t}{k}, \quad r = \binom{t-1}{k-1}, \quad \lambda = \binom{t-2}{k-2}. \quad (38.97)$$

Such a BIB design is called *unreduced*. Since it requires

$$bk = rt = t^{(k)}/(k-1)!$$

observations, it is only useful in practice when k or $(t-k)$ is small. When $k = 2$, (38.93-4) show that in general

$$b = \lambda \binom{t}{2}.$$

This can only be satisfied by an unreduced design (38.97) with $\lambda = 1$.
 When $k = t-1$, (38.97) becomes

$$b = t, \quad r = t-1, \quad \lambda = t-2.$$

This is an example of a *symmetric* BIB design ($b = t, k = r$) which happens here also to be unreduced.

38.52 Since each of t treatments appears r times, and each of b blocks appears k times, in the experiment, it is tempting to interchange the role of treatments and blocks in the design, putting $t' = b, b' = t, r' = k$ and $k' = r$ to obtain a *dual* design. However, this dual will itself be a BIB design if and only if the original BIB design is *symmetric* (cf. Exercise 38.12 for illustrations).

If a BIB design can be resolved into r subsets of blocks, each subset containing each treatment exactly once, the design is called *resolvable*, and each subset is a single *replicate* of the set of treatments. We must then clearly have $t = ck, b = cr$, where c is a positive integer. However, the latter is not alone a sufficient condition for resolvability.

For resolvable BIB designs, (38.96) may be replaced by the stronger inequality, due to Bose (1942),

$$b \geq t + r - 1 \quad (38.98)$$

which is again only a necessary condition for the existence of a resolvable BIB design, for (38.98) actually holds whenever $t = ck, b = cr$ (cf. Exercise 38.13). If and only if the equality holds in (38.98), a resolvable design has k^2/t treatments common to any two blocks in different replicates, and is called an *affine resolvable* BIB design (cf. Exercise 38.18).

38.53 Mann (1949) gives an account of construction methods for BIB designs due to R. C. Bose, whose fundamental series of papers, starting with Bose (1939), is listed in Guérin's (1965) comprehensive bibliography of the subject—these and other methods of construction are summarized by Guérin. Muller (1965) gives a method for obtaining BIB designs from complete sets of orthogonal Latin squares when t is an integral power of a prime number (cf. Exercises 38.10–11 for some simple examples of such constructions given earlier in Fisher and Yates' *Tables*). If t is odd and k does not exceed the smallest prime factor of t , a BIB may always be constructed by the method given in Exercise 38.20. If t is prime, this method is valid for any $k < t$.

38.54 Fisher and Yates' *Tables* give indexes, by the values of k and of r , of all known BIB designs with $r \leq 10$, together with combinatorial methods of obtaining specific designs. Cochran and Cox (1957) give detailed plans for a selection of these designs. C. R. Rao (1961) lists, and gives combinatorial methods for, designs with $11 \leq r \leq 15$ (which are also included and extended in the 6th edition, 1963, of Fisher and Yates' *Tables*) and Sprott (1962) lists designs with $16 \leq r \leq 20$, with references to constructive methods.

Table 38.1 gives, for $t \leq 100$ and $k \leq r \leq 20$, the values of λ for which BIB designs, which are not unreduced, are known to exist. $k = 2$ is omitted from the table since there is then always (cf. 38.51) an unreduced design with $\lambda = 1$. When $k = t-1$, there is always (by 38.51 again) a *symmetric* unreduced design with $\lambda = t-2$, which is

144

THE ADVANCED

$k \backslash t$	6	7	8	9	10	11	12	13	15	16	17	18	19	20	21	22	23	25	26	27	28	29	31	33	35	36	37	39	40	41	43	45	49	52	57	61	64	65	73	81	91				
3	2*	1*																																											
4		3*	3*	2*	6*	3*	1*																																						
5						2*																																							
6																																													
7																																													
8																																													
9																																													
10																																													
11																																													
12																																													
13																																													
14																																													
15																																													
16																																													
17																																													
18																																													
19																																													
20																																													

$k > \frac{1}{2} t$
(see text)

Table 38.1—Index of known BIB designs (excluding unreduced designs) for $t \leq 100$, $k \leq r \leq 20$

The table gives, for any (t, k) , the corresponding value(s) of λ , which may always be increased by any integral multiple by repetition of the whole design. Where more than one value of λ is shown, their integral multiples may be added by adjoining the corresponding repetitions of the different designs.

* Cf. Fisher and Yates' *Tables*: $r \leq 10$

† Cf. C. R. Rao (1961) and Fisher and Yates' *Tables*, 6th edition, 1963: $11 \leq r \leq 15$

‡ Cf. Spott (1962): $16 \leq r \leq 20$

§ Cf. Exercise 38.20

also omitted from our table. Further, we may confine the table to the range $k \leq \frac{1}{2}t$, since (except when $k = t-1$, already discussed) a design with $k' > \frac{1}{2}t$ can always be formed from one with $k < \frac{1}{2}t$ as the *complementary* design, obtained by considering all treatments omitted from each block as a complementary block containing $k' = t - k$ units. We then have, from (38.93-4),

$$\frac{\lambda'}{\lambda} = \frac{(t-k)(t-k-1)}{k(k-1)} > 1.$$

Analysis of BIB designs

38.55 The analysis of an experiment designed in BIB may now be obtained by simple substitution in the results of 38.23-6, which are valid for any block experiment. Using (38.89) and (38.95) in (38.23), we have

$$\begin{aligned}\Omega^{-1} &= r\mathbf{I}_t - \frac{1}{k} \{ (r-\lambda)\mathbf{I}_t + \lambda\mathbf{1}_t\mathbf{1}_t' \} + \frac{r^2}{bk} \mathbf{1}_t\mathbf{1}_t' \\ &= \frac{\lambda t}{k} \left\{ \mathbf{I}_t + \frac{(t-k)}{t^2(k-1)} \mathbf{1}_t\mathbf{1}_t' \right\},\end{aligned}\quad (38.99)$$

on using (38.93-4) to eliminate r and b . (38.99) is of exactly the same form as (38.87), and its inverse, as there, is

$$\Omega = \frac{k}{\lambda t} \left\{ \mathbf{I}_t - \frac{(t-k)}{kt(t-1)} \mathbf{1}_t\mathbf{1}_t' \right\}, \quad (38.100)$$

as the reader may verify directly.

Substituting (38.100) into (38.29) and using (38.34), we find

$$\hat{\tau} = \frac{k}{\lambda t} (\mathbf{T} - \mathbf{nB}/k) + \mathbf{1}_t G/(bk), \quad (38.101)$$

showing that the estimators of the treatment parameters are no longer free of the influence of the block totals—this is as it must be, since different sets of blocks are associated with the various treatments. From the definitions of \mathbf{n} and \mathbf{B} , we see that \mathbf{nB}/k is a $(t \times 1)$ vector whose i th element is the sum of the block averages over all blocks containing the i th treatment. Thus

$$\mathbf{T}_a = \mathbf{T} - \mathbf{nB}/k \quad (38.102)$$

may be called the vector of *adjusted* treatment totals, and is evidently of direct interest. (38.101) becomes

$$\hat{\tau} = \frac{k}{\lambda t} \mathbf{T}_a + \mathbf{1}_t G/(bk). \quad (38.103)$$

(38.32) thus becomes, using (38.103) and (38.4),

$$\hat{\beta} = \mathbf{B}/k - \frac{1}{\lambda t} \mathbf{n}' \mathbf{T}_a - \mathbf{1}_b G/(bk). \quad (38.104)$$

Moreover, the treatment differences SS in the AV table (38.39) is $\mathbf{T}_a' \Omega \mathbf{T}_a$, and using (38.100) and (38.34), this is simply

$$\mathbf{T}_a' \Omega \mathbf{T}_a = \frac{k}{\lambda t} \mathbf{T}_a' \mathbf{T}_a. \quad (38.105)$$

We thus have the following AV table, specialized from (38.39):

AV table for a BIB experiment		D.fr.
Source of variation	SS	
Treatment differences	$\frac{k}{\lambda t} \mathbf{T}'_a \mathbf{T}_a$	$t-1$
Block effects	$\mathbf{B}' \mathbf{B}/k - G^2/(bk)$	$b-1$
Residual	$\mathbf{y}' \mathbf{y} - \frac{k}{\lambda t} \mathbf{T}'_a \mathbf{T}_a - \mathbf{B}' \mathbf{B}/k$	$bk-t-b+1$
General mean	$G^2/(bk)$	1
TOTAL	$\mathbf{y}' \mathbf{y}$	bk

The reader may verify that (38.106) reduces to the randomized blocks AV table (38.52) if we put $k = t$, $\lambda = b$.

Cochran and Cox (1957) and Kempthorne (1952) give detailed instructions for computing the AV in BIB designs, with attention to the simplifications possible in the resolvable, symmetrical, and other special cases. They also take into account the recovery of inter-block information, which we are about to discuss. See also C. R. Rao (1947).

38.56 It might be supposed that the results of 38.55 must complete our discussion of the analysis of BIB designs, but this is not so. The model (38.6) on which all our results are based is a linear model with fixed effects, i.e. we have been carrying out a Model I LS analysis, as in Chapter 35. So far as the treatment parameters are concerned, this is generally realistic, but we have seen in 38.14 that the blocks in an experiment are usually nuisance factors, of no direct interest. The particular blocks used for an experiment are not essential to it. It is not unrealistic, therefore, to consider the block effects as random variables in our analysis. In the terminology of Chapter 36, we are therefore about to consider a *mixed model*, with treatment effects fixed and block effects random. This not unnaturally leads to a different analysis, which is usually called *recovery of inter-block information*. The analysis which follows is not confined to BIB designs, but holds for any block experiment.

Mixed model for the recovery of inter-block information

38.57 We now omit the $(b-1)$ linearly independent block parameters α from the model of 38.21. Instead, we have a random block effect, say $\beta_{(*)}$. If this has zero mean, it will not enter into the expected value of \mathbf{y} , and its variance, say σ_{β}^2 , will be superposed upon the ordinary errors ε_{ij} , which still have zero mean and variance σ^2 . In the notation of 38.16-21, our model is then

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\tau}, \quad (38.107)$$

(*) We break with our usual convention and use a Greek letter for the blocks random variable, since \mathbf{b} and \mathbf{B} are already in use here.

where

$$\mathbf{X}_{(bk \times t)} = \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \\ \vdots \\ \mathbf{t}_b \end{pmatrix}. \quad (38.108)$$

It will be observed that we are still assuming no interactions between the block and treatment effects.

The errors in the model are no longer uncorrelated, since any two observations in the same block share a common value of β . If we write $\sigma_\beta^2/\sigma^2 = \rho$, it is easy to see that the dispersion matrix of \mathbf{y} is

$$\mathbf{V} = \sigma^2 \begin{pmatrix} \mathbf{A} & & \mathbf{0} \\ & \mathbf{A} & \\ \mathbf{0} & & \mathbf{A} \end{pmatrix}, \quad (38.109)$$

where \mathbf{V} has along its leading diagonal b identical matrices

$$\mathbf{A} = \mathbf{I}_k + \rho \mathbf{1}_k \mathbf{1}_k'. \quad (38.110)$$

We suppose initially that ρ is known—we discuss its estimation in 38.62–4.

To estimate τ , we now require the generalized LS estimator, from 19.17 (Vol. 2),

$$\hat{\tau} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y} \quad (38.111)$$

with dispersion matrix

$$\mathbf{V}(\hat{\tau}) = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1}. \quad (38.112)$$

38.58 The inverse of (38.110) is, as at (38.87),

$$\mathbf{A}^{-1} = \mathbf{I}_k - \left(\frac{\rho}{1+k\rho} \right) \mathbf{1}_k \mathbf{1}_k', \quad (38.113)$$

and using (38.108–9) and (38.113), we have

$$\mathbf{X}' \mathbf{V}^{-1} \mathbf{X} = \frac{1}{\sigma^2} \sum_j \mathbf{t}_j' \left\{ \mathbf{I}_k - \left(\frac{\rho}{1+k\rho} \right) \mathbf{1}_k \mathbf{1}_k' \right\} \mathbf{t}_j.$$

Substitution of (38.5) and (38.1) reduces this to

$$\mathbf{X}' \mathbf{V}^{-1} \mathbf{X} = \frac{1}{\sigma^2} \left\{ \text{diag}(\mathbf{r}) - \left(\frac{\rho}{1+k\rho} \right) \sum_j \mathbf{n}_j \mathbf{n}_j' \right\}, \quad (38.114)$$

and since it may be verified from the definitions that

$$\sum_{j=1}^b \mathbf{n}_j \mathbf{n}_j' = \mathbf{nn}', \quad (38.115)$$

(38.114) finally becomes

$$\mathbf{X}' \mathbf{V}^{-1} \mathbf{X} = \frac{1}{\sigma^2} \left\{ \text{diag}(\mathbf{r}) - \left(\frac{\rho}{1+k\rho} \right) \mathbf{nn}' \right\}, \quad (38.116)$$

and the dispersion matrix of the estimators at (38.112) is the inverse of (38.116). Also,

$$\mathbf{X}' \mathbf{V}^{-1} \mathbf{y} = \frac{1}{\sigma^2} \sum_j \mathbf{t}_j' \left\{ \mathbf{I}_k - \left(\frac{\rho}{1+k\rho} \right) \mathbf{1}_k \mathbf{1}_k' \right\} \mathbf{y}_j,$$

and on substituting (38.18) and (38.1), this is

$$\mathbf{X}' \mathbf{V}^{-1} \mathbf{y} = \frac{1}{\sigma^2} \left\{ \mathbf{T} - \left(\frac{\rho}{1+k\rho} \right) \sum_j \mathbf{n}_j \mathbf{B}_j \right\} = \frac{1}{\sigma^2} \left\{ \mathbf{T} - \left(\frac{\rho}{1+k\rho} \right) \mathbf{nB} \right\}. \quad (38.117)$$

Using (38.116-17), (38.111) becomes

$$\hat{\tau} = \left\{ \text{diag}(\mathbf{r}) - \left(\frac{\rho}{1+k\rho} \right) \mathbf{nn}' \right\}^{-1} \left\{ \mathbf{T} - \left(\frac{\rho}{1+k\rho} \right) \mathbf{nB} \right\}. \quad (38.118)$$

38.59 As $\sigma_\beta^2 \rightarrow 0$ ($\rho \rightarrow 0$), (38.118) becomes

$$\lim_{\rho \rightarrow 0} \hat{\tau} = \{\text{diag}(\mathbf{r})\}^{-1} \mathbf{T},$$

and the dispersion matrix of $\hat{\tau}$ is then $\sigma^2 \{\text{diag}(\mathbf{r})\}^{-1}$. These are precisely the results which we should have got if block effects had been completely ignored, and this is as it should be since, as $\sigma_\beta^2 \rightarrow 0$ with $E(\beta) = 0$, the block effects disappear.

At the other extreme, when σ_β^2 (and ρ) $\rightarrow \infty$, we must rewrite (38.118) to avoid the matrix inversion, since now $\frac{\rho}{1+k\rho} \rightarrow \frac{1}{k}$ and $\{\text{diag}(\mathbf{r}) - \mathbf{nn}'/k\}$ is singular, as may be seen by postmultiplying it by $\mathbf{1}_t$ and using (38.3-4). Instead of (38.118), we use

$$\{\text{diag}(\mathbf{r}) - \mathbf{nn}'/k\} \hat{\tau} = \mathbf{T} - \mathbf{nB}/k. \quad (38.119)$$

(38.119) is satisfied by the estimator $\hat{\tau}$ at (38.24) (i.e. (38.29)), provided that the constraint $\mathbf{r}'\boldsymbol{\tau} = G$ is used to make the solution of (38.119) unique. Thus, when block effects are large, the estimators (38.119) tend to coincide with the *intra-block estimators* (38.29). Paradoxical though it may sound, the change to the mixed model only affects the estimators substantially when block effects are small.

It is easy to see that the intra-block estimators remain unbiased in the mixed model, by arguing that since they are so for *any* fixed set of block effects, they must be unconditionally so when the latter become random variables. If the block effects are large, one is then tempted to use the simpler intra-block analysis, since the change in the estimators will, as we have just seen, be small. However, the dispersion matrix of the estimators is now the inverse of (38.116), instead of (38.40).

38.60 It is not quite obvious that in the case of randomized blocks there is no change at all in the estimators obtained by use of the mixed model rather than the fixed-effects one. Exercise 38.14 shows that the estimators (38.118) coincide with the intra-block estimators (38.48) for randomized blocks, but that the dispersion matrix of the estimators is changed in the mixed model, their variances being increased. This is obvious from the fact that any treatment parameter estimator T_i/b is the mean of b independent observations with variance $\sigma^2(1+\rho)$ by (38.109-10).

38.61 Yates' (1939, 1940a) original treatment of the recovery of inter-block information proceeded differently, by observing that *inter-block estimators* of treatment parameters could be obtained from the block totals, that these estimators were uncorrelated with the intra-block estimators, and that the two estimators could therefore be simply weighted to give the smallest attainable variance by this method—the generalized version of this approach is left to the reader as Exercise 38.15. The two different approaches to recovery of inter-block information are not equivalent, although they lead to the same estimator in the BIB case (cf. Exercise 38.16). The reason is that

the weighting of the two components in Exercise 38.15 is determined with reference to the dispersion matrix of the intra-block estimators in the original (fixed-effects) model, while the dispersion matrix of the inter-block estimators is for the mixed model. If the mixed-model dispersion matrix were used for the intra-block estimators also, the two methods would become identical, as is obvious from the fact that $\hat{\tau}$ at (38.118) is a function of \mathbf{T} and \mathbf{B} only.

38.62 Both the MV estimator $\hat{\tau}$ at (38.118) and $\hat{\tau}_I$ in Exercise 38.15 are functions of the variance ratio $\rho = \sigma_{\beta}^2/\sigma^2$, and this must usually itself be estimated by some $\hat{\rho}$ so that $\hat{\tau}(\hat{\rho})$ or $\hat{\tau}_I(\hat{\rho})$ may be used. We first estimate σ_{β}^2 and σ^2 separately and then take the ratio as the estimator of ρ .

To find suitable estimators of σ_{β}^2 and σ^2 , we return to the general analysis of 38.19–24, but we now wish to find an SS attributable to blocks rather than to treatment differences, as in 38.25. We therefore find the Residual SS, say S_2 , when there are no block effects. The difference $S_2 - S_0$ will then be the SS attributable to block effects.

38.63 We put $\hat{\beta} = \mathbf{0}$ in (38.32), and obtain

$$\mathbf{B} = \mathbf{n}' \hat{\tau}_0 \quad (38.120)$$

where $\hat{\tau}_0$ means $(\hat{\tau})_{\hat{\beta}=\mathbf{0}}$. Premultiplying by $\mathbf{1}'_b$ gives, from (38.25) and (38.3),

$$\mathbf{G} = \mathbf{r}' \hat{\tau}_0. \quad (38.121)$$

If we now substitute (38.120–1) into (38.24), we obtain

$$\begin{aligned} \hat{\tau}_0 &= \Omega \{ \mathbf{T} - \mathbf{n} \mathbf{n}' \hat{\tau}_0 / k + \mathbf{r} \mathbf{r}' \hat{\tau}_0 / (bk) \} \\ &= \Omega \{ \mathbf{T} + [\Omega^{-1} - \text{diag}(\mathbf{r})] \hat{\tau}_0 \} \end{aligned} \quad (38.122)$$

on using (38.23). Solving (38.122) for $\hat{\tau}_0$ gives

$$(\hat{\tau})_{\hat{\beta}=\mathbf{0}} = \{ \text{diag}(\mathbf{r}) \}^{-1} \mathbf{T}. \quad (38.123)$$

We then have, in (38.33)

$$S_2 = \mathbf{y}' \mathbf{y} - \mathbf{T}' \{ \text{diag}(\mathbf{r}) \}^{-1} \mathbf{T}, \quad (38.124)$$

and we find

$$\begin{aligned} S_2 - S_0 &= \mathbf{T}' \hat{\tau} + \mathbf{B}' \hat{\beta} - \mathbf{T}' \{ \text{diag}(\mathbf{r}) \}^{-1} \mathbf{T} \\ &= (\mathbf{T} - \mathbf{n} \mathbf{B} / k)' \Omega (\mathbf{T} - \mathbf{n} \mathbf{B} / k) + \mathbf{B}' \mathbf{B} / k - \mathbf{T}' \{ \text{diag}(\mathbf{r}) \}^{-1} \mathbf{T}, \end{aligned} \quad (38.125)$$

using the first row of the AV table (38.39). (38.125) is the required SS attributable to blocks. We thus have the AV table, alternative to (38.39):

Source of variation	SS	D.fr.
Block effects (allowing for treatment differences)	$(\mathbf{T} - \mathbf{n} \mathbf{B} / k)' \Omega (\mathbf{T} - \mathbf{n} \mathbf{B} / k) + \mathbf{B}' \mathbf{B} / k - \mathbf{T}' \{ \text{diag}(\mathbf{r}) \}^{-1} \mathbf{T}$	$b - 1$
Treatment differences (ignoring block effects)	$\mathbf{T}' \{ \text{diag}(\mathbf{r}) \}^{-1} \mathbf{T} - G^2 / (bk)$	$t - 1$
Residual	$\mathbf{y}' \mathbf{y} - \mathbf{T}' \hat{\tau} - \mathbf{B}' \hat{\beta}$	$bk - b - t + 1$
General mean	$G^2 / (bk)$	1
TOTAL	$\mathbf{y}' \mathbf{y}$	bk

The Residual and the General mean rows are unchanged, as they must be, but the remaining SS has been differently apportioned between treatment and block effects because here treatment differences, rather than block effects, are taken into account first. It will be remembered from Example 35.4 that the order is irrelevant only in an orthogonal analysis.

38.64 As usual, we may now estimate σ^2 by the Residual SS S_0 divided by its d.f., since

$$E(S_0) = (bk - b - t + 1)\sigma^2. \quad (38.127)$$

To estimate σ_β^2 , we first observe from (38.124) that the SS due to Blocks (S_B , say) may be written

$$S_2 - S_0 = S_B = \mathbf{y}'\mathbf{y} - S_0 - \mathbf{T}' \{\text{diag}(\mathbf{r})\}^{-1} \mathbf{T},$$

so that, using (38.127),

$$E(S_B) = E(\mathbf{y}'\mathbf{y}) - (bk - b - t + 1)\sigma^2 - E[\mathbf{T}' \{\text{diag}(\mathbf{r})\}^{-1} \mathbf{T}]. \quad (38.128)$$

From the result of Exercise 19.3, we know that if $E(\mathbf{z}) = \mathbf{0}$ and $\mathbf{V}(\mathbf{z}) = \sigma^2 \mathbf{W}$, $E(\mathbf{z}'\mathbf{A}\mathbf{z}) = \sigma^2 \text{tr}(\mathbf{A}\mathbf{W})$. Thus, assuming without loss of generality that $\boldsymbol{\tau} = \mathbf{0}$, we have from (38.109-10)

$$E(\mathbf{y}'\mathbf{y}) = \sigma^2 bk(1 + \rho). \quad (38.129)$$

The model (38.107-10) implies that

$$E(\mathbf{T}) = \{\text{diag}(\mathbf{r})\}\boldsymbol{\tau},$$

$$\mathbf{V}(\mathbf{T}) = \sigma^2[\text{diag}(\mathbf{r}) + \rho \mathbf{nn}'],$$

remembering the properties of \mathbf{nn}' in 38.48. We may thus again apply the result of Exercise 19.3 to obtain

$$\begin{aligned} E[\mathbf{T}' \{\text{diag}(\mathbf{r})\}^{-1} \mathbf{T}] &= \sigma^2 \text{tr} [\{\text{diag}(\mathbf{r})\}^{-1} \{\text{diag}(\mathbf{r}) + \rho \mathbf{nn}'\}] \\ &= \sigma^2 \text{tr} [\mathbf{I}_t + \rho \{\text{diag}(\mathbf{r})\}^{-1} \mathbf{nn}'] \\ &= \sigma^2 \{t + \rho \text{tr} [\{\text{diag}(\mathbf{r})\}^{-1} \mathbf{nn}']\}. \end{aligned}$$

It is easy to verify from the definitions that

$$\text{tr} [\{\text{diag}(\mathbf{r})\}^{-1} \mathbf{nn}'] = t, \quad (38.130)$$

so

$$E[\mathbf{T}' \{\text{diag}(\mathbf{r})\}^{-1} \mathbf{T}] = \sigma^2 t(1 + \rho) \quad (38.131)$$

and (38.129) and (38.131) reduce (38.128) to

$$E(S_B) = (b - 1)\sigma^2 + (bk - t)\rho\sigma^2. \quad (38.132)$$

Thus, from (38.132) and (38.127),

$$E\left\{\frac{S_B - (b - 1)S_0 / (bk - b - t + 1)}{bk - t}\right\} = \rho\sigma^2 \equiv \sigma_\beta^2. \quad (38.133)$$

(38.133) and (38.127) give the required estimators. Their ratio, say $\hat{\rho}$, has no optimum property in this context, where estimation of $\boldsymbol{\tau}$ is of interest. Tocher (1952) gives another, more complicated, estimator of σ_β^2 which, if the errors have zero skewness and kurtosis, is the MV unbiased quadratic estimator. But what is really required is an estimator of ρ whose use allows the treatment parameter estimator (38.118) to remain unbiased.

Graybill and Weeks (1959) and Graybill and Seshadri (1960) show that for

BIB designs, $\hat{\tau}_I$ in Exercise 38.15 remains unbiased when $\hat{\rho}$ is used in it. J. Roy and Shah (1962) show that if the estimator of ρ is a ratio of quadratic forms of a certain type in terms of the latent roots of \mathbf{nn}' , unbiasedness of $\hat{\tau}_I$ is preserved in any incomplete blocks design with $\mathbf{r} = r\mathbf{1}_t$; and that $\hat{\rho}$ is of the required form. Shah (1964) constructs other unbiased estimators than $\tau_I(\hat{\rho})$ in several well-known designs, including the BIB with $t > 5$.

Permutation distributions for BIB designs

38.65 After the general discussion of the mixed model in 38.57–64, we now revert to our earlier fixed-effects model for BIB experiments, and turn our attention to permutation tests for treatment effects.

Ogawa (1963) shows that if there are unit errors (cf. 36.41) the standard F -test for treatments may be justified as an approximation to the permutation distribution if b is large enough and the variances of unit effects within blocks are nearly constant. *A fortiori*, this holds if there are no unit errors and b is large.

If ranks are used within blocks instead of the observations, we may generalize to BIB designs the permutation distribution of the test statistic for treatment effects, discussed in 38.43 and 37.39–41 for randomized blocks. The results, due to Durbin (1951), are given in Exercise 38.17. Van Elteren and Noether (1959) showed that, compared to the usual F -test for treatment effects, Durbin's test using ranks has ARE exactly $k/(k+1)$ times the Wilcoxon ARE (31.115), reducing to $3k/\{\pi(k+1)\}$ in the normal case. It will be seen that the ARE depends on block size, but not upon t .

It is interesting to note that here only the first two moments of the test statistic can be generally obtained, precisely because, as we mentioned at the end of 38.50, the BIB conditions lay down no pattern for the appearance of the treatments in sets of more than two.

Benard and van Elteren (1953) give a large-sample chi-square permutation test for an arbitrary (not necessarily balanced) incomplete blocks design using ranks, repeated as well as missing observations being allowed.

Preference experiments

38.66 BIB designs are of interest in connexion with preference experiments (where measurements of degree of preference are often not possible, but rankings of preferences are). If preferences are to be expressed within b blocks of k objects (treatments) selected from t , the order in which these objects are examined may be important, and it is desirable to arrange the BIB to take this order-effect into account. A simple way is to let the objects be examined in the orders determined by the column positions of a $(t \times t)$ Latin square. If the first k columns of the square are used, they determine order in t blocks of k objects, each of the t possible objects appearing once in each position in the ordering. If $b = ct$, where c is a positive integer, we can obtain complete order-balance by using the first k columns of c $(t \times t)$ Latin squares in this way.

Thus, e.g., the first three columns of (38.71) give a BIB design with $t = 4$, $b = 4$, $k = 3$ and $\lambda = 2$. Each of the letters A, B, C, D occurs once in each column position. An incomplete Latin square, used in this way, is known as a *Youden square* design,

after its discoverer, W. J. Youden. Of course, position-balance may also be important in any (not necessarily preference) BIB experiment, where "position" may stand for any variable whose "nuisance" effect we wish to exclude from the experiment, just as in our original discussion of Latin squares (cf. 38.31 and 38.34).

Paired comparisons

38.67 The particular case $k = 2$ in a BIB design (when, as we saw in 38.51, the design is unreduced) is usually described as a *paired comparisons* design, and is of particular importance in preference experiments. H. A. David (1963) has recently devoted a monograph to methods for paired comparisons, which includes a chapter on appropriate experiment designs. Perhaps the most important of these are the *linked paired-comparison designs* developed by Bose (1956). In these designs, each of t judges compares r pairs of objects (chosen from a total of n objects). Each pair is compared by k judges, and there are exactly λ pairs in common to any two judges. As the notation indicates, there is a correspondence with BIB designs: each judge is a "treatment," each pair of objects a block "containing" k such treatments. There are $b = \frac{1}{2}n(n-1)$ blocks in all. The new feature of the linked paired-comparison designs is that we require each of the n objects to appear equally frequently in the r pairs of each judge, i.e. to appear $2r/n = \alpha$ times. Thus, by (38.94), we have

$$\alpha = k(n-1)/t. \quad (38.134)$$

Because of the additional condition (38.134), the existence of a BIB design does not imply the existence of a linked paired-comparison design, although the latter clearly implies the former. Bose (1956) gives (and David (1963) reproduces) methods for deriving linked paired-comparison designs from BIB designs.

Partially balanced incomplete blocks

38.68 The essential feature of BIB designs (cf. 38.49) is complete symmetry between the treatments, each of which appears r times in all and λ times with any other treatment. This symmetry was a natural consequence of the symmetrical demand for the same precision in all treatment-difference estimators in 38.46-7. While maintaining the condition that each treatment appears r times, we now relax the condition that λ be constant.

Suppose that for each treatment the remaining $(t-1)$ treatments fall into m classes of size t_p , $\sum_{p=1}^m t_p = t-1$. These are called *associate classes*, and any treatment in the p th associate class is called a p th *associate* of the given treatment. We now require that

- (a) all p th associates appear together in the same block λ_p times;
- (b) if A is a p th associate of B , B is a p th associate of A ;
- (c) the number of treatments which are both p th associates of a treatment A , and q th associates of another treatment B , is the same for all l th associates A, B . We write this number as P_{pq}^l .

A design satisfying these conditions is called a *partially balanced incomplete*

blocks (PBIB) design. These designs were first considered by Bose and Nair (1939). Evidently, they contain BIB designs as the special case when $\lambda_p = \lambda$ for all p .

38.69 PBIB designs have many constants: t , r , b and k as before; the m values λ_p and the m values t_p ; and the P_{pq}^l . There are, however, linear relations between these constants which reduce their effective number.

The case of two associate classes ($m = 2$) has been studied in detail by Bose *et al.* (1954) who give tables of all known designs with $r \leq 10$, $3 \leq k \leq 10$. Even in this simplest case, there are five types of *association scheme* (i.e. the scheme which sets out the associate relationships between the treatments), the most important type being the *group divisible* PBIB designs, which themselves consist of three sub-types.

Guérin (1965) gives an extensive summary of the main results on the existence and construction of PBIB designs, with a comprehensive bibliography. Details of the appropriate methods of statistical analysis, including recovery of inter-block information, are given by Bose *et al.* (1954), by Kempthorne (1952), and by C. R. Rao (1947).

Structured treatments: lattice designs

38.70 Throughout our treatment of experiment designs, we have made no assumptions concerning relationships between treatments. Now we suppose that the treatments in the experiment may be meaningfully classified into certain categories. This is the case, e.g., when the treatments are the rc combinations in a two-way cross-classification, the treatments then falling naturally into a two-way table. Block experiments taking account of such classifications are called *lattice designs*, and were introduced by Yates (1936b, 1939, 1940b). They are of particular value when the number of treatments, t , is large, for the table in 38.54 shows how few BIB arrangements are then available.

38.71 Suppose that the t treatments can be meaningfully arranged in a $(l \times k)$ two-way array, so that $t = lk$. We might then use the k treatments in each of the l rows within a single block; and similarly the l treatments in each of the k columns within a single block. We thus obtain a design containing l blocks of k units and k blocks of l units. This is called a *rectangular lattice* design, the name arising from the fact that the treatments may be represented as the points of a lattice in the $(l \times k)$ array. To bring this within the scope of our discussion, which has throughout been limited to blocks of equal size, we must restrict ourselves to the *square lattice* where $k = l$. With two replicates of the treatments as above, it is sometimes called a *simple lattice*; with three replications, a *triple lattice*; with four replications, a *quadruple lattice*, and so on.

The square lattice is not a balanced design in the sense that a BIB is, for although every treatment appears $r = 2$ times, it is not true that every treatment appears equally frequently in the same block with every other treatment. The reader will see at once that the frequency of joint appearance λ_{ii} is unity for treatments in the same row or column of the $(k \times k)$ array, and zero for all other pairs of treatments.

One can evidently generalize the two-dimensional $(k \times k)$ array of treatments to a p -dimensional array containing k^p treatments. We then have k^{p-1} blocks of k units

in a single replication of the treatments. Such p -dimensional lattice designs are again not balanced. The cubic lattice ($p = 3$) is important in some applications.

38.72 To avoid the more complicated analysis attending upon lack of balance, we may construct *balanced lattice* designs by replicating the simple lattice arrangements just discussed. In the case $t = k^2 = 9$, the 3×3 array

$$\begin{array}{ccc} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{array} \quad (38.135)$$

yields the simple square lattice design as described in 38.71 with six blocks

$$\begin{array}{c|c|c} 1 & 4 & 7 \\ \hline 2 & 5 & 8 \\ \hline 3 & 6 & 9 \end{array} \parallel \begin{array}{c|c|c} 1 & 2 & 3 \\ \hline 4 & 5 & 6 \\ \hline 7 & 8 & 9 \end{array} \quad (38.136)$$

arranged in two complete replications of the treatments. If we now add two further replications

$$\begin{array}{c|c|c} 1 & 2 & 3 \\ \hline 5 & 6 & 4 \\ \hline 9 & 7 & 8 \end{array} \parallel \begin{array}{c|c|c} 1 & 2 & 3 \\ \hline 6 & 4 & 5 \\ \hline 8 & 9 & 7 \end{array} \quad (38.137)$$

the *ensemble* of (38.136-7) is fully balanced, as the reader may verify. In fact, it forms a BIB with $t = 9$, $k = 3$, $\lambda = 1$, $b = 12$, $r = 4$. The four complete replications in (38.136-7) form a *set of lattice* squares, because they can be derived as in 38.71 from (38.135) and the further array

$$\begin{array}{ccc} 1 & 5 & 9 \\ 6 & 7 & 2 \\ 8 & 3 & 4 \end{array} \quad (38.138)$$

These designs are more valuable when $k = t^{\frac{1}{2}}$ is odd, only $\frac{1}{2}(k+1)$ squares then being required, as in our example. When k is even, $k+1$ squares are needed to form a set.

38.73 Details of the theory and analysis of lattice designs, into which we have no space to enter here, are given by Kempthorne (1952) and by Cochran and Cox (1957). Their importance for our exposition is that they have led us to consider a set of treatments which are "structured," at least to the extent of being arranged in categories. If we pursue this line, and turn our attention to treatments which are combinations of underlying elements, we are led into new territory.

Factorial experiments

38.74 If the treatments in an experiment consist of all possible combinations of a set of underlying factors, it is called a *factorial experiment*. Such experiments are formally the same as the (complete) cross-classifications discussed in detail in 35.15-33 and 35.40-4 (although we did not use the terminology of experiments in earlier chapters because the analyses given there are also applicable to non-experimental situations). Each defining variable of the classification (i.e. the row-variable, column-variable, etc.) is called a *factor* in the experiment. Each value which a factor can take,

which defines an individual cell of the marginal classification by that factor, is called a *level* of the factor. Thus what we previously called a $(r \times c)$ cross-classification may be described in this context as a factorial experiment with two factors, one at r levels and the other at c levels; and a $(r \times c \times l)$ cross-classification is a factorial experiment with three factors, at r , c and l levels respectively. More concisely, these two examples could be described as a $(r \times c)$ factorial and as a $(r \times c \times l)$ factorial respectively; if $r = c$, the former would be called a (r^2) factorial, and so on.

38.75 If a factorial experiment is carried out in a randomized blocks design, we naturally wish to subdivide the treatments SS (with $(t-1)$ d.fr.) into component parts for main effects and interactions. Thus, e.g., in a $(r \times c)$ factorial experiment with one observation per cell, we should have $t = rc$, and the $(rc-1)$ d.fr. for treatment differences are to be resolved into $r-1$, $c-1$ and $(r-1)(c-1)$ as in Example 35.3. More generally, the treatment differences SS are to be subdivided into components for all the main effects and the different-order interactions, as in Chapter 35. We call this subdivision an *AV for treatments*.

It will be seen from (38.52) that the treatment differences SS is $\frac{1}{b}\{T'T - G^2/t\}$, so that $T'T/b$ now plays the same role, for a factorial experiment in randomized blocks, as $y'y$ did in Chapter 35. Thus an AV for treatments may be carried out upon the treatment totals T_i by exactly the methods of Chapter 35, reading t as n . It is necessary only to remember to divide all the component SS by b , this divisor arising, of course, because each T_i is the sum of b observations.

38.76 From Exercise 38.6 it will be seen that the same simple AV for treatments in terms of the T_i may be carried out for factorial experiments in Latin square designs, the divisor here being t (instead of b) which is again the number of observations of which T_i is the sum. The same rule holds for the generalization of Exercise 38.6 to Graeco-Latin and higher-order orthogonal square designs.

Confounding

38.77 So far, the subdivision of the treatment differences SS has been simplicity itself, because in 38.75-6 we have considered only factorial experiments (themselves of simple structure) in the simplest block designs. We now remove both these limitations, and return to the consideration of experiments upon a set of related treatments in the general block design.

A glance at the treatment differences SS in the general block experiment AV table (38.39) will show the reader that we cannot now expect the simplicity of 38.75-6 to persist, for the allocation of treatments to blocks vitally affects the treatment differences SS through **B**. This remains true even for the BIB table (38.106), remembering the definition of T_a at (38.102).

A little thought will convince the reader that this is inevitable, for if the treatments do not all appear in the same blocks, their effects must become entangled with those of the blocks themselves, and balance (in the BIB sense) is not enough to preclude this.

Indeed, a new problem now arises, for there is evidently a danger that important components of the treatment differences SS (e.g. main effects and first-order interactions) cannot be estimated because they are inextricably entangled with the block effects. They are then said to be *confounded* with blocks.

38.78 Rather than confine ourselves to the SS in the AV table, we consider quite generally the estimation of p linear functions of the treatment parameters, say $C\tau$, where C is a $(p \times t)$ matrix of known coefficients. In particular, we are interested in contrasts in the parameters, where (cf. 35.58) the elements in each row of C sum to zero. Contrasts, it will be recalled, include simple differences, and also interactions, between the treatment parameters.

Inspection of our model for block experiments at (38.12-14) shows that

$$C\tau \equiv (C \mid 0)\theta, \quad (38.139)$$

where 0 is a $(p \times (b-1))$ matrix of zeros which serves to annihilate the block parameters α .

Now in order that a vector Ly be unbiased in estimating (38.139), it is necessary and sufficient by (19.19) that

$$LX = (C \mid 0)$$

i.e. using (38.13),

$$\left. \begin{aligned} L \begin{pmatrix} t_1 \\ \vdots \\ t_b \end{pmatrix} &= C, \\ L \begin{pmatrix} 1_k u'_1 \\ \vdots \\ 1_k u'_b \end{pmatrix} &= 0. \end{aligned} \right\} \quad (38.140)$$

Thus, if we can find a matrix L , of order $(p \times bk)$, which satisfies (38.140), there will be no confounding of the p linear functions $C\tau$.

38.79 The equations (38.140) impose $p(t+b-1)$ conditions upon the pbk elements of L , and whatever p may be, these can certainly be satisfied if $t+b-1 \leq bk$, i.e. if $b \geq (t-1)/(k-1)$. The equations may also be satisfiable for some values of p if $b < (t-1)/(k-1)$, for the conditions upon L are not necessarily independent ones—this depends on the structure of the block design. The case $t = bk$ (where the experiment consists of a single replication of the set of treatments) falls into this category, for $t/k < (t-1)/(k-1)$ for $t > k$, while $t = k$ is trivial in this context.

We therefore see that, if there are enough blocks, we can always, if we wish, avoid confounding any set of linear functions of the treatment parameters. This is intuitively obvious from the fact that if certain functions are confounded in a given set of blocks, we may deliberately add a further set of blocks in which they are not confounded. This is so in particular where each set of blocks is a replication of the treatments. In

the literature, functions confounded in part, but not all, of the experiment are said to be *partially confounded*.

38.80 We have so far discussed confounding as though it were an evil to be avoided. It is undoubtedly a nuisance to be unable to estimate some functions of the treatment parameters, and even in the case of partial confounding there are computational complications which may be irksome, while naturally the precision of the estimators of the partially confounded functions must be reduced.

However, confounding also has its positive aspect in factorial experiments. It will be remembered from **35.44** that the higher-order interactions are commonly found to be of little practical value. They are therefore often deliberately confounded with the blocks in an experiment, the consequence being that their SS and d.fr. appear as part of the Residual. Of course, we may carry out precisely the same merging process in an unconfounded analysis, as indicated at **35.44**. The point here is that within a given framework of experiment (t , k and b), it may not be possible to estimate all the desired linear functions of the treatment parameters, namely the main effects and interactions. If some of these must be confounded, it is in general advantageous to start with the highest-order interactions and confound as few of the main effects and first-order interactions as possible.

To this end, Fisher (1942) proved, using Abelian group theory, that in a factorial experiment with $2^m - 1$ factors each at two levels (the $2^{2^m - 1}$ factorial), no main effect or first-order interaction need be confounded in a single replication of the treatments, provided that $k \geq 2^m$, i.e. if block size exceeds the number of factors. He later (Fisher (1945)) extended his treatment to factors with p^r levels, where p is a prime (cf. also Mann (1949)). Kempthorne (1952) gives a very detailed treatment of the subject, including factors with different numbers of levels. Cochran and Cox (1957) discuss the applications, with detailed plans of confounded block arrangements. Yates (1937) gives many examples, with applications in agricultural experimentation, while Davies *et al.* (1954) give examples in industrial experiments.

38.81 One of the important applications of confounding is in *fractionally replicated* factorial experiments, where certain interactions are assumed to be zero in order that the remaining main effects and interactions may be estimated by using only some of the blocks of a confounded block design. The assumptions are essential to enable the analysis to distinguish between otherwise indistinguishable effects, which are called *aliases*. The theory is treated by Kempthorne (1952), and some discussion appears in Cochran and Cox (1957). Davies *et al.* (1954) discuss fractionally replicated designs in their application to industrial experiments, where they are often useful when large numbers of factors are to be tested.

38.82 Many other confounded factorial designs are now available to the experimenter. *Split-plot* designs confound main effects by assigning every level of a factor to the units in a block, an arrangement which is sometimes convenient and even necessary for the practical conduct of the experiment. Other, more elaborate, forms

of confounding in factorial experiments (quasi-Latin squares, plaid squares) are described and analysed in the specialized books to which we have already referred.

Sequences of experiments: evolutionary operation

38.83 The book by Quenouille (1953a) lays particular emphasis on the problems of planning and analysis of long-term experiments and groups of experiments (see also Cochran and Cox (1957) and Kempthorne (1952)).

A different but related field which has been opened up recently is the use of a sequence of experiments to find the *optimum combination* of factors, i.e. to maximize the yield (or some other quality) of the end-product of a process for fixed cost, or equivalently to minimize cost for a fixed yield. The method is to fit a *response surface* to the experimental points by LS, and to move the area of experimentation along a path of steepest ascent until it is near a stationary point, which is then explored to investigate its character. Experiment designs with special properties of symmetry, called *rotatable* designs, have been developed for the exploration of response surfaces. For the theory of this method of *evolutionary operation* and the associated rotatable designs, the reader should consult Box and Wilson (1951), Box and Hunter (1957), Bose and Carter (1959), Gardiner *et al.* (1959), Bose and Draper (1959), Draper (1960a, b, c, 1961), and Box and Behnken (1960). Less theoretical expositions are given in the final chapter of Davies *et al.* (1954) and by Box (1957).

Evolutionary operation methods are not properly sequential methods—they have no well-defined overall probabilistic properties, and have been criticized on this score—but there can be no doubt of their practical importance.

Regression designs

38.84 We end this chapter with an account of the design of experiments whose object is to carry out a regression analysis.

As we mentioned in 38.2, Example 28.4 dealt with the problem of designing a simple linear regression experiment. We found there that we could minimize the sampling variances of the LS estimators of the two regression parameters by making half of the observations as far below the origin as possible, and the other half the same distance above the origin. We remarked there that this corresponds to the fact that a straight line is most efficiently “fixed” by its end-points.

Consider now the more general problem of allocating values to the regressor x when the expected value of y is a polynomial in x . We have treated the theory of polynomial regression in 28.16–20, taking the values of x for granted; now we ask how to choose values of x so that the parameters of the polynomial regression equation are optimally estimated.

38.85 We assume that the values of x are to be allocated in a fixed interval, which without loss of generality may be called $(-1, +1)$. The intuitive argument above extends to the polynomial case, for we know that a polynomial of degree k can be “fixed” by $(k+1)$ points. Moreover, one still expects in this polynomial case that one of the points “ought” to be at each end of the interval. This intuitive argument is asymptotically valid for the general polynomial regression: Kiefer (1959) shows

that at most $(k+1)$ distinct values of x are required, of which at most $(k-1)$ are interior to the interval, provided that we ignore the analytical complications due to n being an integer, which disappear as $n \rightarrow \infty$. Characteristically, however, the exact theory, taking these complications into account, is not so simple—intuitive arguments cannot be expected to hold here.

38.86 Now consider the choice of the $(k+1)$ distinct observation points $x_1 < x_2 < \dots < x_{k+1}$ to minimize the generalized variance of the estimators of the parameters. In this polynomial case we have

$$y = X\theta + \epsilon$$

where the matrix has the form

$$\underset{(n \times (k+1))}{X} = \begin{pmatrix} 1_{n_1} & 1_{n_1} x_1 & 1_{n_1} x_1^2 & \dots & 1_{n_1} x_1^k \\ 1_{n_2} & 1_{n_2} x_2 & 1_{n_2} x_2^2 & \dots & 1_{n_2} x_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1_{n_{k+1}} & 1_{n_{k+1}} x_{k+1} & 1_{n_{k+1}} x_{k+1}^2 & \dots & 1_{n_{k+1}} x_{k+1}^k \end{pmatrix} \quad (38.141)$$

where n_i observations are taken at the point x_i , $i = 1, 2, \dots, k+1$ and $\sum_{i=1}^{k+1} n_i = n$.

Thus the dispersion matrix of $\hat{\theta}$ is

$$V(\hat{\theta}) = \sigma^2 (X'X)^{-1} = \sigma^2 (Z'NZ)^{-1} \quad (38.142)$$

where

$$\underset{((k+1) \times (k+1))}{Z} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_{k+1} \\ x_1^2 & x_2^2 & \dots & x_{k+1}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^k & x_2^k & \dots & x_{k+1}^k \end{pmatrix} \quad (38.143)$$

and

$$\underset{((k+1) \times (k+1))}{N} = \begin{pmatrix} n_1 & & & 0 \\ & n_2 & & \\ & & \ddots & \\ 0 & & & n_{k+1} \end{pmatrix}. \quad (38.144)$$

We therefore see that the effect of having $(k+1)$ observation points and $(k+1)$ parameters is to make $X'X$ a product of three square matrices. Hence

$$\left| \frac{1}{\sigma^2} V(\hat{\theta}) \right|^{-1} = \frac{\sigma^{2(k+1)}}{|V(\hat{\theta})|} = |X'X| = |Z'| |N| |Z| = |N| |Z|^2, \quad (38.145)$$

and if the generalized variance is to be minimized, (38.145) is to be maximized. We carry out this maximization in two stages. First,

$$|N| = \prod_{i=1}^{k+1} n_i$$

is at once seen to be maximized for choice of the n_i when they are all equal, whatever $|Z|^2$ may be. The latter alone remains to be maximized for choice of the x_i .

38.87 Now the reader may verify that

$$|Z| = \prod_{\substack{i,j=1 \\ i < j}}^{k+1} (x_i - x_j) \quad (38.146)$$

so that

$$|Z|^2 = \prod_{i < j} (x_i - x_j)^2.$$

It is obvious at once from the form of (38.146), a product of squares, that it can always be increased by moving the extreme observation points to the ends of the interval. What is more, (38.146) when maximized will always be larger for a larger interval than for a smaller. Thus we see that we should always observe over the largest possible interval and locate one observation point at each end of that interval, which we shall continue to refer to as $(-1, +1)$. At each observation point, an equal number $n/(k+1)$ of observations is to be made.

38.88 We may now solve (38.146) *ad hoc* for the smaller values of k . For $k = 1$, the linear case, nothing further is needed, for the result of 38.87 has already confirmed Example 28.4. For the quadratic case, we have to locate x_2 , with $x_1 = -1$ and $x_3 = +1$, to maximize (38.146), which is now simply

$$|Z|^2 = 4(1 - x_2^2)^2.$$

This is maximized at $x_2 = 0$, and this must obviously be the other observation point, from consideration of symmetry.

In the cubic case, we explicitly use the symmetry to reduce the problem to locating x_3 and $x_2 = -x_3$. (38.146) is now

$$|Z|^2 = 16x_3^2(1 - x_3^2)^4$$

which is maximized when $x_3^2 = \frac{1}{5}$.

In the quartic case, symmetry locates x_3 at zero, and we require only x_4 and $x_2 = -x_4$. (38.146) becomes

$$|Z|^2 = 16x_4^6(1 - x_4^2)^4,$$

maximized when $x_4^2 = \frac{3}{7}$.

These results, and the next two, which are as many as are needed in practice, are summarized in the following table:

Degree of polynomial, k	Observation points in $(-1, +1)$ at which $n/(k+1)$ observations should be made
1	± 1
2	$\pm 1, 0$
3	$\pm 1, \pm 0.4472$
4	$\pm 1, \pm 0.6547, 0$
5	$\pm 1, \pm 0.7651, \pm 0.2852$
6	$\pm 1, \pm 0.8302, \pm 0.4689, 0$

(38.147)

38.89 Hoel (1958) showed that the optimum observation points which maximize (38.146) are expressible in terms of the Legendre polynomials which are defined in 30.37. Guest (1958) considered a different criterion of optimality, the minimization of the maximum variance of the fitted polynomial at any point in the interval, which led to the same optimum values; he showed that the optimum values are given explicitly by the zeros of the derivative of the k th order polynomial. The same criterion of optimality had been considered in a paper by K. Smith (1918), who first calculated the values (38.147). This was apparently the first design problem to be solved in detail, and it is all the more surprising that the paper was more or less forgotten for forty years. Smith's paper also contains a series of charts comparing the variance of the fitted polynomial throughout the interval when the observations are made (a) by the optimum method; (b) by the method of uniform spacing of observations, which is better in the centre of the interval, but much worse at the extremes; and (c) by method (b) with an additional group of observations at each end of the interval, which removes the worst effects of purely uniform spacing. The advantage of method (c), of course, is that it does not presuppose any knowledge of k , and enables the experimenter to investigate its value from the observations, whereas the optimum method of allocation cannot be used to investigate a higher value of k —this is precisely the point which we made in Example 28.4 for the linear case. It seems wise, in any case, to use the values in (38.147) corresponding to the highest value of k which the experimenter would be willing to consider.

K. Smith (1918) goes on to consider the effect of heteroscedasticity of errors on the optimum allocation. Hoel (1958) considers some special cases of correlated observations.

38.90 Other criteria of optimality have also been used. Hoel and Levine (1964) consider the allocation of observations in polynomial regression to minimize the variance of the fitted polynomial at a specified point outside the interval of observation $(-1, +1)$, and it transpires that this optimum allocation also minimizes the maximum variance over an interval $(-1, x)$ if a certain condition is satisfied by the value x . Gaylor and Sweeny (1965) consider minimizing the maximum variance, and an average variance, over any interval (arbitrarily related to the interval of observation) for the linear case only. H. A. David and Arens (1959) consider using two observation points in the linear case to minimize expected mean-square-error or maximum expected squared-error, the latter differing from minimizing maximum variance since the possibility is allowed that the true degree of the polynomial may be 2 rather than 1. A very general paper by Kiefer and Wolfowitz (1959), not confined to the polynomial case, uses game-theoretic results to compute optimum allocations using a number of criteria (see also the summary in Kiefer (1959)). Hoel (1965a) applies these methods to obtain optimum allocations for two-dimensional polynomial and trigonometric regressions. Hoel (1965b) finds the designs in univariate polynomial regression which minimize the variance of the fitted value for x "extrapolated" into an interval lying immediately between two intervals in which observations are made, and also obtains corresponding designs in bivariate polynomial regression, as well as for ordinary extrapolation in the bivariate case.

THE ADVANCED THEORY OF STATISTICS

EXERCISES

38.1 In 38.25, verify the identity of the two expressions (38.38) for the SS attributable to treatment differences in a block experiment.

38.2 By generalizing the argument of 38.28, show that if we divide the treatment parameter estimators into groups and require zero correlation between members of different groups, each block must contain the same number of treatments from a given group, and thus the experiment may be resolved into a set of independent sub-experiments with smaller blocks.
(Tocher, 1952)

38.3 In 38.28, show that if a block experiment is designed to make the variance of each treatment parameter estimator equal to its value if there are no block effects, $\sigma^2 \{\text{diag}(\mathbf{r})\}^{-1}$, this requires that $\mathbf{M} = \mathbf{0}$, and hence leads to the randomized blocks design with incidence matrix given by (38.46).
(Tocher, 1952)

38.4 Verify the simplified formulae (38.47-51) for randomized blocks designs.

38.5 Verify the formulae in 38.32-3 for the LS analysis of blocks classified in a two-way table, and construct the AV table as at (38.39).

38.6 From the general results of 38.33, show that the AV table for the Latin square design in 38.35 is:

Source of variation	SS	D.fr.
Treatment differences	$\mathbf{T}'\mathbf{T}/t - G^2/t^2$	$t-1$
Rows	$\mathbf{R}'\mathbf{R}/t - G^2/t^2$	$t-1$
Columns	$\mathbf{C}'\mathbf{C}/t - G^2/t^2$	$t-1$
Residual	$\mathbf{y}'\mathbf{y} - (\mathbf{T}'\mathbf{T} + \mathbf{R}'\mathbf{R} + \mathbf{C}'\mathbf{C})/t + 2G^2/t^2$	$(t-1)(t-2)$
General mean	G^2/t^2	1
TOTAL	$\mathbf{y}'\mathbf{y}$	t^2

Generalize this to Graeco-Latin and higher-order orthogonal square designs.

38.7 In 38.39, show that no more than $(t-1)$ Latin squares of order t can be mutually orthogonal.

38.8 Verify that the inverse of (38.79) is

$$\Omega^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}(\mathbf{w} \mid \mathbf{1}_t) \begin{pmatrix} 1 + \mathbf{1}'_t \mathbf{D}^{-1} \mathbf{w} & -\mathbf{1}'_t \mathbf{D}^{-1} \mathbf{1}_t \\ -\mathbf{w}' \mathbf{D}^{-1} \mathbf{w} & 1 + \mathbf{1}'_t \mathbf{D}^{-1} \mathbf{w} \end{pmatrix} \begin{pmatrix} \mathbf{1}'_t \\ \mathbf{w}' \end{pmatrix} \mathbf{D}^{-1} / \Delta,$$

where Δ is the determinant of the 2×2 matrix above. Hence show through (38.23) that (38.85) holds for any \mathbf{D} .

38.9 Show that if \mathbf{D} is determined so that every treatment difference is estimated with the same loss of efficiency (compared to the situation where there are no block effects to eliminate),
(Tocher, 1952)

(38.85) leads to (38.95), the BIB design equation. If the loss of efficiency is zero, show that this reduces to the randomized blocks design (cf. Exercise 38.3).
(Tocher, 1952)

38.10 In a superposed complete set of orthogonal Latin squares of order T (e.g. (38.76) for $T = 4$), each of the T^2 cells of the arrangement has $(T+1)$ "references," identifying its row, its column, and its position in each of the $(T-1)$ "alphabets" forming the complete set, and each reference can take T distinct values. Show that if each reference in turn is used to allocate each of the cells to one of a set of T blocks, we obtain a BIB design with

$$t = T^2, \quad b = T(T+1), \quad k = T, \quad r = T+1, \quad \lambda = 1.$$

Verify that in the case of (38.76), the resulting BIB design is:

Block number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Treatments in block	1	5	9	13	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	2	6	10	14	5	6	7	8	6	5	8	7	7	8	5	6	8	7	6	5
	3	7	11	15	9	10	11	12	11	12	9	10	12	11	10	9	10	9	12	11
	4	8	12	16	13	14	15	16	16	15	14	13	14	13	16	15	15	16	13	14
Obtained from	Rows				Columns				Roman letters				Greek letters				Numerals			

38.11 In Exercise 38.10, show that if we augment each of the T blocks obtained from a particular reference by a further treatment, where a different such treatment is used for each distinct reference, and finally add a further single block containing all the $(T+1)$ further treatments, we obtain a symmetric BIB design with $t = b = T^2 + T + 1$, $k = r = T + 1$, $\lambda = 1$. Verify this augmented design for the design derived from (38.76) in Exercise 38.10.

By considering the case $T = 6$ (cf. 38.39) show that satisfaction of (38.93-5) is not sufficient for the existence of a BIB design.

38.12 In Exercise 38.10, show that the dual design obtained by putting $t' = b$, $b' = t$, $k' = r$ and $r' = k$ has $\lambda < 1$ and hence cannot be a BIB design. In Exercise 38.11, show that the dual of the augmented design derived from (38.76) is:

Block number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Treatments in block	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4	1	5	9	13	17
	5	6	7	8	5	6	7	8	5	6	7	8	5	6	7	8	2	6	10	14	18
	9	10	11	12	10	9	12	11	11	12	9	10	12	11	10	9	3	7	11	15	19
	13	14	15	16	15	16	13	14	16	15	14	13	14	13	16	15	4	8	12	16	20
	17	18	19	20	20	19	18	17	18	17	20	19	19	20	17	18	21	21	21	21	21

Verify that this is a BIB design.

38.13 Show from (38.93) that for any BIB design,

$$r/(t-1) = (r-\lambda)/(t-k),$$

and with $t = ck$, $b = cr$ as above (38.98), use (38.93-4) to show that

$$\frac{r(c-1)}{t-1} = \frac{r-\lambda}{k} = \frac{b-r}{t-1} = r - c\lambda = I,$$

where I is a positive integer. Hence establish (38.98).

(V. N. Murty, 1961)

38.14 Using (38.45-6), show that for randomized blocks the estimator (38.118) is identical with the intra-block estimator (38.29), both reducing to (38.48), so that use of the mixed model does not change the estimator in this case. Show that the dispersion matrix of the estimators, (38.112), is $\frac{\sigma^2}{b}\{I_t + \rho I_t I_t'\}$, differing from the intra-block result ((38.40) and (38.47)), the variances in particular being increased.

38.15 In 38.57, use (38.1) to show that

$$E(B) = n'\tau$$

and

$$V(B) = k(1+k\rho)\sigma^2 I_b,$$

so that we may estimate the treatment parameters unbiasedly from the block totals by

$$\hat{\tau}_B = (nn')^{-1}nB,$$

with

$$V(\hat{\tau}_B) = k(1+k\rho)\sigma^2 (nn')^{-1},$$

provided that nn' is non-singular, which implies $b \geq t$.

Show that $\hat{\tau}_B$ is uncorrelated with $\hat{\tau}$ defined at (38.29), and hence use the generalized LS method of 19.17 (Vol. 2) and the dispersion matrix (38.40) to show that the linear combination of them with smallest variances is

$$\hat{\tau}_I = \left\{ \text{diag}(r) - \left(\frac{\rho}{1+k\rho} \right) nn' + rr'/(bk) \right\}^{-1} \left\{ T - \left(\frac{\rho}{1+k\rho} \right) nB + rG/(bk) \right\}.$$

Show that $\hat{\tau}_I$ reduces to (38.29) when $\rho \rightarrow \infty$.

(cf. Tocher, 1952)

38.16 Show that $\hat{\tau}_I$ of Exercise 38.15 coincides with the MV estimator at (38.118) if and only if

$$\tau' \left\{ \text{diag}(r) - \left(\frac{\rho}{1+k\rho} \right) nn' \right\}^{-1} \left\{ T - \left(\frac{\rho}{1+k\rho} \right) nB \right\} = G.$$

Verify that the equality holds when the design equation is of the BIB form (38.95).

(Sprott (1956) obtained a similar result, but instead of $\hat{\tau}_I$ used an estimator combining $\hat{\tau}$ and $\hat{\tau}_B$ in Exercise 38.15 with weights reciprocal to the variances of their elements.)

38.17 In a BIB design, the observations within each block are replaced by their ranks 1, 2, ..., k. If T_i is the total of the ranks thus allotted to the i th treatment, and

$$S_t = \sum_{i=1}^t (T_i - T.)^2 \equiv \sum T_i^2 - t\{\frac{1}{2}r(k+1)\}^2,$$

with maximum value $S_{\max} = \lambda^2 t(t^2 - 1)/12$, show by extending 37.40-1 and Exercise 37.14 that $W = S_t/S_{\max}$ has

$$E(W) = \frac{(k+1)}{\lambda(t+1)},$$

$$\text{var } W = \frac{2(k+1)^2}{\lambda^2 r t(t+1)^2} \left(r-1 + \frac{\lambda-1}{k-1} \right)$$

exactly, and that, approximately,

$$\left\{ \frac{\lambda(t+1)}{k+1} - 1 \right\} W/(1-W)$$

has the variance-ratio distribution with d.fr.

$$v_1 = t - 1 - \frac{2}{r}$$

$$v_2 = (r - 1)v_1,$$

reducing to the results of Exercise 37.14 when $k = t$, $\lambda = r$, and the BIB becomes a randomized blocks design.

(Durbin, 1951)

38.18 For any BIB design resolvable into r replicates of the t treatments, show that the number of treatments common to two blocks in different replicates has mean equal to k^2/t and sum of squares about this mean proportional to $(b - t - r + 1)$. Hence show that (38.98) holds, and that if and only if the equality holds in (38.98), there are exactly the same number (i.e. k^2/t) of treatments common to any two blocks in different replicates.

(Bose, 1942)

38.19 Verify the observation points given for $k = 5$ and $k = 6$ in the table (38.147).

(K. Smith, 1918)

38.20 If t is odd, p is its smallest prime factor, and $k \leq p$, show that a BIB may always be constructed by the following *equal-differences* method. Label the treatments $0, 1, 2, \dots, t-1$ and construct $\frac{1}{2}(t-1)$ initial blocks containing the treatments $[0, 1, 2, \dots, k-1]; [0, 2, 4, \dots, 2(k-1)]; \dots; [0, \frac{1}{2}(t-1), t-1, \dots, \frac{1}{2}(k-1)(t-1)]$, where every number in the blocks is calculated as a residue (mod t). From each initial block form a new block by adding to each of its treatments an integer r ; this is done for every such integer $1 \leq r \leq t-1$. We thus obtain a set of t blocks (one for each member of the residue class mod t) for each initial block. Show that the resulting BIB has $b = \frac{1}{2}t(t-1)$, $r = \frac{1}{2}k(t-1)$ and $\lambda = \frac{1}{2}k(k-1)$.

(Gassner, 1965)

CHAPTER 39

SAMPLE SURVEY THEORY: DESIGNS

The estimation of means of finite populations

39.1 At the beginning of our discussion of random sampling in 9.1-4, Vol. 1, we pointed out that the application of probability theory to a sampling procedure requires only that it should specify the chances of selection of all possible samples; it is by no means necessary that the procedure should be simple random sampling, which gives every possible sample from the population the same chance of selection. We went on to distinguish between sampling without replacement from a finite population, which necessarily involves dependence between successively selected members of the population, and sampling with replacement, which removes that dependence.

For almost the whole of these three volumes, we have been (and shall be) concerned only with simple random sampling. In this and the following chapter, however, we are specifically interested in other forms of random sampling. These arise, in practice, when there is a problem of sampling a finite population, i.e., a population with a finite number (which we shall always denote by N) of members. Even in sampling with replacement, we now have a new feature arising from the finiteness of the population: every individual in the population is *recognizable*, so that if a value appears twice in the sample, we can know whether it is the same individual appearing twice, or two different individuals with the same value.

The direct interest of sample survey theory, as this branch of the subject has come to be called, is almost entirely in the estimation of means (or, equivalently, totals) of the variables being studied. The theory which we shall study is nevertheless more general than this, since the mean of any function of a variable (e.g. of its square) may be treated by the same method. We shall study the estimation of variances and other constants of the population only in so far as this is necessary to throw light upon our central concern, the estimation of means. Results for proportions are always derivable from those for means by specializing the variable to take values 0 or 1 only.

We are thus entering a rather narrow area of statistical theory, but it is an area which has been intensively cultivated, and this on grounds of its practical importance rather than of its mathematical attractiveness. The large journal literature has in recent years been summarized and supplemented by several books, notably those of Cochran (1963), Hansen *et al.* (1953) and Yates (1960). The last-mentioned book contains extensive bibliographies of the theoretical and applied work on survey sampling which have been brought up to date in new editions since its first publication in 1949. M. N. Murthy (1963) reviews recent theoretical developments.

We shall have to confine our discussion to the theoretical aspects of sample surveys, and our aim will be to display them in the context of general statistical theory. We cannot hope that all the results of importance will be elegantly displayed, but we shall try to minimize the inherent cumbrousness of the subject.

Random sampling with equal probabilities without replacement

39.2 We wish to estimate the mean $\mu^{(*)}$ of a variable y in a finite population with N members, from a sample of n members drawn at random without replacement, using equal probabilities of selection. It is perhaps not quite obvious that (in the absence of any knowledge of the form of the population) the MV estimator of μ is the sample mean m . We now use the Least Squares theory of Chapter 19 to demonstrate this.

Consider the model

$$\underset{(n \times 1)}{\mathbf{y}} = \underset{(n \times 1)(1 \times 1)}{\mathbf{1}\mu} + \underset{(n \times 1)}{\boldsymbol{\epsilon}} \quad (39.1)$$

where $\mathbf{1}$ is a vector of units. The errors ϵ_i are the deviations of the observations from the population mean, and they are not uncorrelated, because the drawings are not independent. By the symmetry of the situation, the covariance, say $\rho\mu_2$, between any pair of observations in the sample is the same. It is this symmetry which leads us to expect the sample mean to be the best estimator of μ . The dispersion matrix of the errors is

$$\mathbf{V}(\boldsymbol{\epsilon}) = \mu_2 \begin{pmatrix} 1 & \rho & . & . & . & . & \rho \\ \rho & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & \rho \\ \rho & . & . & . & . & \rho & 1 \end{pmatrix} = \mu_2 \mathbf{V}. \quad (39.2)$$

By 19.17 and Exercise 19.5, the MV unbiased linear estimator of μ is the LS estimator

$$\hat{\mu} = (\mathbf{1}' \mathbf{V}^{-1} \mathbf{1})^{-1} \mathbf{1}' \mathbf{V}^{-1} \mathbf{y} \quad (39.3)$$

with variance

$$\mathbf{V}(\hat{\mu}) = \mu_2 (\mathbf{1}' \mathbf{V}^{-1} \mathbf{1})^{-1}. \quad (39.4)$$

To use (39.3-4), we must evaluate \mathbf{V}^{-1} . As may be verified by multiplication with \mathbf{V} , this is

$$\mathbf{V}^{-1} = \frac{1}{(1-\rho)\{1+(n-1)\rho\}} \begin{pmatrix} \{1+(n-2)\rho\} & -\rho & . & . & . & -\rho \\ -\rho & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & -\rho \\ -\rho & . & . & . & -\rho & \{1+(n-2)\rho\} \end{pmatrix} \quad (39.5)$$

Hence

$$\begin{aligned} \mathbf{1}' \mathbf{V}^{-1} &= \{1+(n-1)\rho\}^{-1} \mathbf{1}', \\ \mathbf{1}' \mathbf{V}^{-1} \mathbf{1} &= n\{1+(n-1)\rho\}^{-1} \end{aligned}$$

(*) For convenience, we write μ (without suffix) for the mean μ'_1 of the population in this and the next chapter only. The suffixed moments μ_r will denote the higher central moments as usual. The corresponding sample moments are written (again in these chapters only) m and m_r , $r > 2$.

and (39.3-4) becomes

$$\hat{\mu} = m,$$

$$V(\hat{\mu}) = \frac{\mu_2}{n} \{1 + (n-1)\rho\}. \quad (39.6)$$

So far, we have not had to evaluate the correlation, ρ . We have delayed this deliberately, since the whole of this section remains valid whatever ρ may be. In our present application, we may easily evaluate ρ by noting that when $n = N$, $V(\hat{\mu})$ reduces to zero, since the whole population is sampled. Thus, from (39.7), $\rho = -\frac{1}{N-1}$, when $n = N$. But clearly, the correlation between a pair of sample values is independent of sample size. Thus, quite generally for random sampling without replacement,

$$\rho = -\frac{1}{N-1} \quad (39.8)$$

and (39.7) becomes

$$V(\hat{\mu}) = \frac{\mu_2}{n} \frac{N-n}{N-1}. \quad (39.9)$$

The moments of the sample mean and their unbiased estimators

39.3 In fact, (39.9) and also the third and fourth central moments of the distribution of m have already been derived in Example 12.8 using combinatorial techniques. We there found it algebraically convenient to work in terms of sample and population k -statistics, k_r and K_r , and throughout this chapter we shall do the same. In fact, we shall *redefine* the population "variance" as

$$\sigma^2 = K_2 = \frac{N}{N-1} \mu_2$$

and the sample "variance" as

$$s^2 = k_2 = \frac{n}{n-1} m_2;$$

we retain the name "second moment" for μ_2 and m_2 . With this notation, we rewrite the results of Example 12.8 (which include (39.9)) as(*)

$$\left. \begin{aligned} E(m) &= \mu, \\ V(m) &= \sigma^2 \left(\frac{1}{n} - \frac{1}{N} \right), \\ E(m-\mu)^3 &= K_3 \left\{ \left(\frac{1}{n^2} - \frac{1}{N^2} \right) - \frac{3}{N} \left(\frac{1}{n} - \frac{1}{N} \right) \right\}, \\ E(m-\mu)^4 &= K_4 \left\{ \left(\frac{1}{n^3} - \frac{1}{N^3} \right) - \frac{4}{N} \left(\frac{1}{n^2} - \frac{1}{N^2} \right) \right. \\ &\quad \left. + \frac{6}{N^2} \left(\frac{1}{n} - \frac{1}{N} \right) \right\} + 3 \frac{(N-1)}{(N+1)} \left(\sigma^2 - \frac{K_4}{N} \right) \left\{ \left(\frac{1}{n^2} - \frac{1}{N^2} \right) - \frac{2}{N} \left(\frac{1}{n} - \frac{1}{N} \right) \right\}. \end{aligned} \right\} \quad (39.10)$$

(*) We now discard the suffix N to the expectation operator E ; it is always present by implication in this chapter and the next.

We recall from (12.109) that

$$E(k_x) = K_x, \quad (39.11)$$

so that we may substitute s^2 for σ^2 and k_3 for K_3 in (39.10) to obtain the unbiased estimators of the central moments of m ,

$$\left. \begin{aligned} \hat{V}(m) &= s^2 \left(\frac{1}{n} - \frac{1}{N} \right), \\ \{E(\hat{m} - \mu)^3\} &= k_3 \left\{ \left(\frac{1}{n^2} - \frac{1}{N^2} \right) - \frac{3}{N} \left(\frac{1}{n} - \frac{1}{N} \right) \right\}. \end{aligned} \right\} \quad (39.12)$$

The unbiased estimator of the fourth moment is slightly more complicated, since it is not linear in the K_r . We require an unbiased estimator of $\sigma^4 = K_2^2$. Exercise 12.11 gives at once the result

$$E \left\{ \frac{k_2^2 - \frac{(N-n)(Nn-n-N-1)}{n(n-1)N(N+1)} k_4}{\frac{2(N-n)}{(n-1)(N+1)} + 1} \right\} = K_2^2,$$

which reduces to

$$E \left\{ \left(\frac{n-1}{n+1} \right) \left(\frac{N+1}{N-1} \right) k_2^2 - \frac{(N-n)(Nn-n-N-1)}{n(n+1)N(N-1)} k_4 \right\} = K_2^2. \quad (39.13)$$

Substitution of the random variable in braces in (39.13) for σ^4 , and k_4 for K_4 , in the last equation of (39.10) gives an unbiased estimator of the fourth moment of m .

39.4 There is nothing to prevent us, in any given case, from estimating the first four moments of m as indicated in 39.3, and then fitting a Pearson distribution to obtain an estimate of the sampling distribution; the situation here is absolutely analogous to that of Maximum Likelihood estimation in 18.20, Vol. 2. However, here as there the process of fitting a small-sample distribution by moments is rarely carried out. This is less due to laziness than to the fortunate fact that the Central Limit theorem makes the labour unnecessary for the sample sizes encountered in practice. Although we shall not prove the limiting normality of m , there is a new point worth considering in connexion with the nature of the limiting process. We cannot simply let n tend to infinity, since it cannot exceed N . Thus Madow (1948), who established a Central Limit result in this case, allowed both n and N to increase, subject only to n/N remaining bounded away from 1. It is easy to verify from (39.10) that the skewness and kurtosis coefficients of m tend to the normal values under this limiting process. Hájek (1960) gives a necessary and sufficient condition for the limiting normality of m .

We may thus apply the standard error techniques, as described in 9.26-9, Vol. 1, to the distribution of m , and carry out tests of hypothetical values μ_0 of the population mean, or set confidence limits for μ , in the ordinary way. It is only necessary that n be large enough. If n/N is small, we may effectively proceed as for simple random sampling.

Sufficiency in sample survey theory: individuals recognizable

39.5 In 17.31, Vol. 2, we developed the concept of sufficiency. A sufficient statistic yields all the information in the sample concerning the parameter. Now, in sample survey theory, we do not specify the form of the population distribution, but following Basu (1958) and Pathak (1964c) we may take as parameter the N -dimensional vector of the values of the variable y in the population. Any sample of n observations (whether selected with or without replacement) will then yield information as to n or fewer of the N elements of the parameter-vector. Any two samples (selected by the same sampling scheme) which contain the same set of $d(\leq n)$ population members yield the same information about the parameter, irrespective of whether these d members appear with different frequencies in the two samples; they are therefore called *equivalent samples*. The set of all possible samples obtainable by a given sampling scheme can be partitioned into subsets in many ways. If each subset of a partition consists of equivalent samples, that partition is called *sufficient*. If some of the subsets of a sufficient partition can be merged and another sufficient partition thus formed, the latter is called a smaller sufficient partition. If there were a sufficient partition to which every other sufficient partition could be reduced by such a merging process, it would be a minimal sufficient partition (cf. 23.16, Vol. 2).

If we define any (vector) statistic (i.e. any function of the sample values of y) t , we induce a partition of the set of all possible samples, each subset consisting of all samples with a particular value of t . If the partition thus induced by t is sufficient, t itself is a sufficient statistic, since its value characterizes subsets of equivalent samples. The conditional distribution of any other statistic, given t , will tell us nothing further about (i.e. be free of) the parameter.

The general theory of sufficient statistics in Chapters 17 and 23 now applies. In particular, the Rao-Blackwell result of 17.35 states that, given any unbiased estimator of a function of the parameter, we can improve upon it by using instead its conditional expectation given a sufficient statistic.

Simple as this result is, it has some unexpected consequences. Consider the simplest case of random sampling with replacement from a finite population of N members, to estimate the population mean. The intuitive estimator is the sample mean, m . However, this can be improved upon, for in general the n sample members will include some individuals selected more than once, and, as we have seen, it is the $d \leq n$ *distinct* individuals in the sample which form the basis for equivalent samples. Thus the vector statistic t , consisting of the values $y^{(1)}, y^{(2)}, \dots, y^{(d)}$ attached to these distinct individuals in the sample, is a sufficient statistic. The conditional expectation of m given t is evidently the mean of the distinct individuals in the sample, say m_d , and this will have smaller variance than m . Raj and Khamis (1958) give the reduction in variance explicitly—see Exercise 39.1. There, as quite often, sample survey estimators improved by the Rao-Blackwell method have rather complicated variances to evaluate.

As a general rule, in sampling with replacement, it will always improve precision of estimation if only the *distinct* selected individuals are used in any estimation process, rather than all the individuals selected. The variance of the latter estimator is, however, usually easier to estimate, and is always an upper bound to that of the former estimator,

so may be used conservatively in place of it. Unless n/N is large, the reduction in variance will rarely be sufficiently large to justify the extra labour required to ascertain its extent.

Pathak (1961a, b, 1962a, b, 1964a, b, c) has carried out a series of investigations of the application of the Rao-Blackwell method to various problems in sample survey theory.

Sampling without replacement with unequal probabilities

39.6 We now generalize random sampling without replacement by allowing the probabilities of selection to differ between individuals and from drawing to drawing. Let ${}_{(r)}p_i$ be the probability that the i th individual (with value y_i) is selected at the r th drawing, $\sum_{i=1}^N {}_{(r)}p_i = 1$; i ranges from 1 to N and r from 1 to n . Now let

$$\pi_i = \sum_{r=1}^n {}_{(r)}p_i, \quad (39.14)$$

$$\pi_{ij} = \sum_{\substack{r,s=1 \\ r \neq s}}^n {}_{(r)}p_i {}_{(s)}p_j, \quad (39.15)$$

the later probability on the right of (39.15) being taken as conditional upon the earlier event being realized. From their definitions, π_i is the overall probability that y_i is selected for the sample of size n , and π_{ij} is the joint probability that both y_i and y_j ($i \neq j$) are selected for the sample. Clearly, the complete set of ${}_{(r)}p_i$, which we call the *selection scheme*, determines the π_{ij} and π_i ; but the same set of π_{ij} and π_i may be associated with different selection schemes. It is usually a difficult matter to find values of the ${}_{(r)}p_i$ to achieve desired values of π_{ij} and π_i , but in selection schemes such as those in Exercises 39.5-6 the connecting equations can be solved numerically. Fellegi (1963) gives a recursive method of making the ${}_{(r)}p_i$ equal for every drawing. Equal-probabilities sampling without replacement, which we have already considered, has

$$\pi_i = \frac{n}{N}, \quad \pi_{ij} = \frac{n(n-1)}{N(N-1)}.$$

In general the reader may verify from (39.14-15) that

$$\sum_{i=1}^N \pi_i = n; \quad \sum_{j=1}^N \pi_{ij} = (n-1)\pi_i, \quad j \neq i; \quad \sum_{\substack{i,j=1 \\ i \neq j}}^N \pi_{ij} = n(n-1). \quad (39.16)$$

We now follow the notational convention that sample observations are labelled y_1, y_2, \dots, y_n in the order in which they are drawn. This will not generally coincide with the order of labelling of the population y_1, y_2, \dots, y_N . This means that, e.g., y_3 in the sample is not y_3 in the population, but in the interests of simplicity we shall retain this notation when we are considering symmetric functions of the sample values.

In taking expectations, we consider the sample as a whole, and take $nm = \sum_{i=1}^n y_i$

as the random variable. There are $\binom{N}{n}$ possible samples, and we suppose these to be listed in some order and labelled $S_1, S_2, \dots, S_r, \dots, S_{\binom{N}{n}}$. By definition,

$$nE(m) = E\left(\sum_{i=1}^n y_i\right) = \sum_{r=1}^{\binom{N}{n}} \text{Prob}\{S_r\} \left(\sum_{i=1}^n y_i\right)_r.$$

The $\binom{N}{n}$ sets of n values in the summations may be reassembled into N sets of $\binom{N-1}{n-1}$ values, corresponding to the $\binom{N-1}{n-1}$ ways in which each of the N individuals in the population can enter the sample. Thus

$$nE(m) = \sum_{i=1}^N \left[\sum_{r=1}^{\binom{N-1}{n-1}} \text{Prob}\{S_r\} \right]_i y_i = \sum_{i=1}^N \pi_i y_i = \tau, \quad (39.17)$$

say. Thus, as we should expect, the sample mean is not generally unbiased for the population mean. We also have

$$\begin{aligned} n^2 V(m) &= V\left(\sum_{i=1}^n y_i\right) = E\left\{\left(\sum_{i=1}^n y_i\right)^2\right\} - \tau^2 \\ &= E\left(\sum_{i=1}^n y_i^2\right) + E\left(\sum_{\substack{i,j=1 \\ i \neq j}}^n y_i y_j\right) - \tau^2. \end{aligned}$$

The first expectation is evaluated exactly as at (39.17), and is equal to $\sum_{i=1}^N \pi_i y_i^2$. In the second expectation, there are $\binom{N}{n}$ sets of $n(n-1)$ values, which we reassemble in $N(N-1)$ sets of $\binom{N-2}{n-2}$ values, corresponding to the $\binom{N-2}{n-2}$ ways in which each of the $N(N-1)$ pairs of individuals in the population can enter the sample. Thus

$$\begin{aligned} E\left(\sum_{\substack{i,j=1 \\ i \neq j}}^n y_i y_j\right) &= \sum_{r=1}^{\binom{N}{n}} \text{Prob}\{S_r\} \left(\sum_{\substack{i,j=1 \\ i \neq j}}^n y_i y_j\right)_r = \sum_{\substack{i,j=1 \\ i \neq j}}^N \left[\sum_{r=1}^{\binom{N-2}{n-2}} \text{Prob}\{S_r\} \right]_{i,j} y_i y_j \\ &= \sum_{\substack{i,j=1 \\ i \neq j}}^N \pi_{ij} y_i y_j \end{aligned}$$

and therefore

$$\begin{aligned} V\left(\sum_{i=1}^n y_i\right) &= n^2 V(m) = \sum_{i=1}^N \pi_i y_i^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^N \pi_{ij} y_i y_j - \tau^2 \\ &= \sum_{i=1}^N \pi_i (1 - \pi_i) y_i^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^N (\pi_{ij} - \pi_i \pi_j) y_i y_j. \end{aligned} \quad (39.18)$$

39.7 The expectations which we evaluated to obtain (39.17-18) are special instances of the general results, for any function g of the observations

$$\left. \begin{aligned} E\left\{\sum_{i=1}^n g(y_i)\right\} &= \sum_{i=1}^N \pi_i g(y_i), \\ E\left\{\sum_{\substack{i,j=1 \\ i \neq j}}^n g(y_i, y_j)\right\} &= \sum_{\substack{i,j=1 \\ i \neq j}}^N \pi_{ij} g(y_i, y_j) \end{aligned} \right\} \quad (39.19)$$

which require no further proof, since the argument which we used remains unchanged when y_i and $y_i y_j$ are replaced by arbitrary functions.

We may now obtain an unbiased linear estimator of μ . If the same weight w_i is to be attached to an individual value y_i in the population whenever it is selected, we must have

$$E\left\{\sum_{i=1}^n w_i y_i\right\} = \mu,$$

and hence, by (39.19) with $g(y_i) = w_i y_i$,

$$\sum_{i=1}^N \pi_i w_i y_i = \mu = \frac{1}{N} \sum_{i=1}^N y_i.$$

This must hold identically in μ , so we equate coefficients of y_i and find $w_i = 1/(N\pi_i)$ and thus

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^n \frac{y_i}{\pi_i} \quad (39.20)$$

is the only unbiased estimator of this form. Further, the variance of $\hat{\mu}$ is simply

(39.18) with $\frac{y_i}{N\pi_i}$ replacing y_i . Thus

$$N^2 V(\hat{\mu}) = \sum_{i=1}^N \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^N \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j. \quad (39.21)$$

39.8 We can obtain an unbiased estimator of (39.21) by use of (39.19). By inspection, such an estimator is

$$N^2 \hat{V}_1(\hat{\mu}) = \sum_{i=1}^n \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^n \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j \pi_{ij}} y_i y_j, \quad (39.22)$$

proposed by Horvitz and Thompson (1952), who first formulated the problem as in 39.6. On the other hand, we may use (39.16) to write (39.21) in the identical form

$$N^2 V(\hat{\mu}) = \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \quad (39.23)$$

Thus, by (39.19) a second unbiased estimator of (39.21) is

$$N^2 \hat{V}_2(\hat{\mu}) = \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^n \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2, \quad (39.24)$$

which was proposed by Yates and Grundy (1953).

If we could choose each π_i proportional to y_i , the variance (39.23) and its estimator (39.24) would be identically zero. (Exercise 39.3 shows that this is not necessarily so for the alternative estimator of variance (39.22).) In practice, we can at best approximate to this situation if we have knowledge of a variable highly correlated with y . Fellegi (1963) and J. N. K. Rao (1963) discuss methods of achieving desired values of the π_i and their effects on the sampling variance of the estimator.

Hájek (1964) investigates the asymptotic normality, variance and estimated variance of (39.20) for *rejective sampling*, which is sampling with unequal probabilities with replacement in which the whole sample is rejected as soon as any individual is selected a second time.

It is not at first clear which of these alternative estimators of sampling variance is preferable. It is easily shown that the estimator (39.22) can take negative values (cf. Exercise 39.3), but it is not so obvious that (39.24) can also take negative values—Exercise 39.4 shows that this is the case. It is nevertheless true that in two selection schemes used in practice, (39.24) is never negative—Exercises 39.5–6 give details—and this is not so for (39.22). There seems no doubt that (39.24) is generally preferable to (39.22), since the latter is, so to say, more likely to take negative values, and for an estimator of variance (with necessarily positive expectation) this implies a larger sampling variance.

It seems likely that there is no estimator of the sampling variance (39.21) which is non-negative definite for all selection schemes, but so far as we know this has not been demonstrated.

39.9 However, by adopting a different linear estimator from (39.20), we can put ourselves in the position of always having a non-negative estimator of the sampling variance of our estimator. In order to do this, we must no longer confine ourselves to linear functions of the form $\sum_{i=1}^n w_i y_i$ discussed when defining $\hat{\mu}$ at (39.20), for we saw there that $\hat{\mu}$ is the *unique* such linear function which is unbiased.

In constructing linear estimators of the population mean from a set of sample values, the coefficients attached to each sample value may be made to depend on:

- the individual population member whose selection for the sample yields that value;
- the drawing at which the sample value was selected;
- the whole set of population members selected for the sample, rather than the individual member as in (a).

The coefficients may depend on any one, two, or all three of (a), (b), (c), and there are thus seven general classes of linear estimator, the last class including all the others. They are discussed and analysed by Godambe (1955, 1965) and by Koop (1963), who show that we cannot find a minimum variance unbiased estimator in this most general class, a proposition which is intuitively plausible from the consideration that, given knowledge of the population values, we can always construct a linear unbiased estimator with zero variance, but that this estimator must clearly depend on the population values themselves. An example of such an estimator is (39.20), as we see at Exercise 39.3. Essentially, the absence of an optimum linear estimator here is due to the fact

that the individuals in the population are recognizable and the probabilities of selection are at choice.

We have so far considered only coefficients depending on (a) alone. We now consider an estimator whose coefficients depend on all three of (a), (b), (c).

39.10 Let $y_{(r)}$ now denote the individual selected at the r th drawing. Let $p_{(r)}$ be the conditional probability of its selection at that drawing, given that it has not been selected previously, and define

$$z_1 = y_{(1)}/p_{(1)}, \quad z_u = \sum_{r=1}^{u-1} y_{(r)} + y_{(u)}/p_{(u)}, \quad u = 2, 3, \dots, n. \quad (39.25)$$

Each of the nz_u is unbiased for $N\mu$, since

$$E(z_1) = \sum_{i=1}^N y_i$$

by (39.19), and for $u \geq 2$,

$$\begin{aligned} E(z_u) &= E\left[\sum_{r=1}^{u-1} y_{(r)} + E\left\{\frac{y_{(u)}}{p_{(u)}} \mid y_{(1)}, \dots, y_{(u-1)}\right\}\right] \\ &= \left[\sum_{r=1}^{u-1} y_{(r)} + \left\{N\mu - \sum_{r=1}^{u-1} y_{(r)}\right\}\right] = E\{N\mu\} = N\mu. \end{aligned}$$

Thus any linear function

$$z = \sum_{u=1}^n c_u z_u, \quad \sum_{u=1}^n c_u = 1 \quad (39.26)$$

will be unbiased for $N\mu$. The most symmetrical such function is the mean

$$\bar{z} = \frac{1}{n} \sum_{u=1}^n z_u = \frac{1}{n} \left\{ \sum_{u=1}^n \frac{y_{(u)}}{p_{(u)}} + \sum_{u=2}^n \sum_{r=1}^{u-1} y_{(r)} \right\}. \quad (39.27)$$

(39.27) does not reduce to Nm when probabilities of selection are equal—cf. Exercise 39.10. The variance of \bar{z} is

$$V(\bar{z}) = \frac{1}{n^2} \left\{ \sum_{u=1}^n \text{var } z_u + \sum_{\substack{u, v=1 \\ u \neq v}}^n \text{cov}(z_u, z_v) \right\}. \quad (39.28)$$

By evaluating the covariance of z_u and z_v in two stages, the first for z_u fixed, and the second allowing z_u to vary, it follows at once that $\text{cov}(z_u, z_v) = 0$, $u \neq v$. Hence

$$E(z_u z_v) = E(z_u) E(z_v) = N^2 \mu^2. \quad (39.29)$$

Thus we only have to evaluate the variances in (39.28). In general, $\text{var } z_u$ is cumbersome to evaluate (see Exercise 39.9), but it is easy to estimate $V(\bar{z})$, for

$$V(\bar{z}) = E(\bar{z}^2) - N^2 \mu^2$$

so that, using (39.29), an unbiased estimator of $V(\bar{z})$ is $\bar{z}^2 - z_u z_v$ for any $u \neq v$. If we average this estimator over all $\frac{1}{2}n(n-1)$ distinct pairs u, v , we obtain the estimator

$$\hat{V}(\bar{z}) = \bar{z}^2 - \frac{1}{n(n-1)} \sum_{u \neq v} z_u z_v, \quad (39.30)$$

which is identical with

$$\hat{V}(\bar{z}) = \frac{1}{n(n-1)} \sum_{u=1}^n (z_u - \bar{z})^2 \geq 0. \quad (39.31)$$

(39.31) is of the form $\frac{s^2}{n}$, where s^2 is the second k -statistic of the observed z_u . This approach, due to Raj (1956), therefore reduces the problem of sampling the y_i (with unequal probabilities and without replacement) to sampling the z_u and calculating their mean and estimating its sampling variance exactly as though they were sampled with equal probabilities and with replacement. This is a special case of a general result given as Exercise 39.12.

39.11 We have to decide which of the estimators (39.20) and (39.27) to use. Little is known in general of their relative efficiencies, but Raj (1956) reports two sampling experiments with $n = 2$ which favour (39.27) fairly strongly.

However, an estimator with variance never greater than (39.27) can be obtained by applying the Rao-Blackwell method of improving an estimator—see 39.5. We calculate (39.27) (which we now relabel $\bar{z}_{(s)}$ to emphasize its dependence on the order of drawing the sample) for each of the $n!$ possible orders in which the observed sample could have been drawn, and then average these $n!$ values using the relative probabilities of the $n!$ orderings as weights, thus obtaining an improved estimator \bar{z}_s . The estimator of variance at (39.31) can be treated in exactly the same way to obtain an improved estimator of $V(\bar{z}_{(s)})$. These results, due to M. N. Murthy (1957), are direct consequences of the Rao-Blackwell result of 17.35 and are given in Exercises 39.30–1. The improved estimator \bar{z}_s has not been shown to possess a non-negative estimator of variance for all selection schemes, as $\bar{z}_{(s)}$ has, but Murthy showed that this is so for $n = 2$. Moreover, in Raj's (1956) sampling experiments already mentioned, M. N. Murthy (1957) confirmed that \bar{z}_s is more efficient than $\bar{z}_{(s)}$ and that an unbiased estimator of its sampling variance fluctuates less than (39.31) does for $\bar{z}_{(s)}$. Finally, the improved estimator of $V(\bar{z}_{(s)})$ also achieved worthwhile gains in efficiency over (39.31).

There are thus strong theoretical reasons for using the improved estimator \bar{z}_s . However, computational problems become formidable as n increases, since $n!$ different sample orderings have to be considered, and even $\bar{z}_{(s)}$ as originally defined at (39.27) requires the computation of n conditional probabilities, which may require considerable labour when population size N and sample size n are large. When the selection scheme is simplified, as in Exercise 39.31, by keeping probabilities proportional at all drawings, \bar{z}_s takes a simpler form, but still seems formidable to compute.

Unequal probabilities in sampling with replacement

39.12 When sampling is with replacement, drawings are independent, and we now have the possibility that any y_i is selected more than once. We define π_i and π_{ij} by (39.14–15) as before, and now require the further definition

$$\pi_{ii} = \sum_{\substack{r, s=1 \\ r \neq s}}^n (r) p_i (s) p_i.$$

π_{ij} is now the probability that y_i and y_j are selected once or more for the sample, and π_{ii} the probability that y_i is selected at least twice. (39.16) is replaced by

$$\sum_{i=1}^N \pi_i = n; \quad \sum_{j=1}^N \pi_{ij} = (n-1)\pi_i; \quad \sum_{i=1}^N \sum_{j=1}^N \pi_{ij} = n(n-1), \quad (39.32)$$

which holds for any i, j without restriction. Similarly (39.19) is replaced by

$$\left. \begin{aligned} E\left\{\sum_{i=1}^n g(y_i)\right\} &= \sum_{i=1}^N \pi_i g(y_i), \\ E\left\{\sum_{\substack{i,j=1 \\ i \neq j}}^n g(y_i, y_j)\right\} &= \sum_{i,j=1}^N \pi_{ij} g(y_i, y_j), \end{aligned} \right\} \quad (39.33)$$

i and j being unrestricted in the final double summation in (39.33). In (39.18) and (39.21) the π_{ij} may now have its suffixes equal, but the term $\pi_i \pi_j$ in the double summation must still have different suffixes, as the reader should verify. If we now use (39.32) instead of (39.16), we find that (39.23) holds without change.

In the particular case when the probabilities of selection are the same at each drawing, so that we may write them p_i without a prefix, the theory simplifies, for we now have $\pi_i = np_i$ and $\pi_{ij} = n(n-1)p_i p_j$, and the estimator (39.20) becomes

$$\hat{\mu} = \frac{1}{Nn} \sum_{i=1}^n \frac{y_i}{p_i}, \quad (39.34)$$

while (39.23) becomes

$$N^2 V(\hat{\mu}) = \frac{1}{2n} \sum_{i,j=1}^N p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2, \quad (39.35)$$

and using (39.33), the unbiased estimator of (39.35) is

$$N^2 \hat{V}_2(\hat{\mu}) = \frac{1}{2n^2(n-1)} \sum_{i,j=1}^n \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2.$$

Using (2.27), Vol. 1, this is

$$N^2 \hat{V}_2(\hat{\mu}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{p_i} - \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i} \right)^2 = \frac{s_{y/p}^2}{n}, \quad (39.36)$$

where s^2 is the second sample k -statistic, defined as in 39.3, of y_i/p_i . The simplicity of the result (39.36) springs from the fact that the $y_i/(Np_i)$ are unbiased estimators of μ , uncorrelated because sampling is with replacement, so that Exercise 39.12 applies here.

Sample designs: stratification

39.13 We saw in 39.8 that we cannot hope in practice to find a selection scheme which will reduce the sampling variance (39.23) to its attainable minimum of zero. As indicated there, however, we may have available some general (though possibly rather imprecise) information about the variable y , or some other variables correlated with it, which enables us to improve considerably on simple random sampling. The aim of a *sample design* (i.e. a choice of the π_{ij} and hence the π_i) is to reduce estimation variance as much as possible. (Later, we shall modify this statement to take into account the varying costs of different sampling procedures.) If, as usual, we are dealing with several variables simultaneously, we can at best find some compromise set of π_{ij} which will be effective in producing small variances for all the estimators we shall use. It is this need for compromise which lends point to the consideration of various classes of selection schemes which have the aim of variance-reduction in mind.

39.14 First, consider the form of (39.23) when the π_i are fixed and only the π_{ij} are at choice (within the system of constraints imposed by (39.16)). (39.23) is a sum of $\frac{1}{2}N(N-1)$ terms, each of which is a coefficient $(\pi_i\pi_j - \pi_{ij})$ times a non-negative quantity $\left(\frac{y_i - y_j}{\pi_i - \pi_j}\right)^2$ which is now fixed by the values of the π_i . If the values y_i are unknown, it is not possible to say which pattern of π_{ij} will be optimum, as we have already seen, but irrespective of the values of the y_i , we see that whenever we fix a π_{ij} equal to $\pi_i\pi_j$, the coefficient will be zero. Now, we saw in 39.6 that in ordinary equal-probabilities sampling without replacement, we have $\pi_i = \frac{n}{N}$, $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$ for all $i \neq j$, whence $\pi_i\pi_j - \pi_{ij} = \frac{n(N-n)}{N^2(N-1)} > 0$, and every one of the $\frac{1}{2}N(N-1)$ terms in (39.23) makes a positive contribution to the variance in this case. We are now in a position to see that if we put every $\pi_i = n/N$ and some $\pi_{ij} = \pi_i\pi_j = n^2/N^2$, we are bound to reduce the contribution to the variance of our estimator from those pairs i, j . But because of (39.16) the π_{ij} sum to $n(n-1)$, and the slight increase in some π_{ij} from $\frac{n(n-1)}{N(N-1)}$ to $\frac{n^2}{N^2}$ will be offset by a decrease in other π_{ij} , increasing the corresponding contributions to the variance. However, if these latter reduced π_{ij} are associated with smaller values of $|y_i - y_j|$, while the increased π_{ij} are associated with larger values of $|y_i - y_j|$, we should expect a net overall reduction in sampling variance. Moreover, since the increased π_{ij} are only slightly increased, the compensating reduction in the other π_{ij} need only be small, especially if there are no fewer reduced than increased π_{ij} .

We thus have arrived at a rather imprecise principle for improving sampling variance with the π_i all equal: increase the π_{ij} from $\frac{n(n-1)}{N(N-1)}$ to $\frac{n^2}{N^2}$ wherever $|y_i - y_j|$ is large, and decrease the π_{ij} wherever $|y_i - y_j|$ is small. It will make our principle clearer if we realize that when $\pi_{ij} = \pi_i\pi_j$, y_i and y_j must be selected by *independent* processes. Thus we are investigating a procedure in which the selection scheme is broken up into two or more sub-schemes, operated quite independently of each other. Reverting to the notation of 39.6, we may express this in terms of the selection scheme as follows:

$$\left. \begin{array}{l} \text{For} \\ \text{for} \\ \text{for} \end{array} \right\} \begin{array}{l} 1 \leq i \leq N_1, \\ N_1 < i \leq N_1 + N_2, \\ \vdots \\ \sum_{l=1}^{k-1} N_l < i < N, \end{array} \quad \left. \begin{array}{l} (r)p_i \left\{ \begin{array}{l} > 0 \text{ for } 1 \leq r \leq n_1, \\ = 0 \text{ otherwise;} \end{array} \right. \\ (r)p_i \left\{ \begin{array}{l} > 0 \text{ for } n_1 < r \leq n_1 + n_2, \\ = 0 \text{ otherwise;} \end{array} \right. \\ \vdots \\ (r)p_i \left\{ \begin{array}{l} > 0 \text{ for } \sum_{l=1}^{k-1} n_l < r \leq n, \\ = 0 \text{ otherwise.} \end{array} \right. \end{array} \right\} \quad (39.37)$$

In (39.37), the N population members are split into k groups containing N_i members, $\sum_{i=1}^k N_i = N$: the sample is similarly split into k corresponding samples of sizes n_i , $\sum_{i=1}^k n_i = n$. Each sub-population is independently sampled.

Our principle now tells us to choose the y_i as members of the groups so that the members of different groups are as different as possible (for the zero coefficients in (39.23) will come from pairs in different groups) while the members of any one group are as alike as possible (for these are the pairs whose π_{ij} will be decreased).

Population sub-groups, each of which is to be sampled independently and the results combined to estimate overall population parameters, are called *strata*, the groups thus being identified metaphorically with geological layers in the population. For the detailed study of *stratified sampling*, as our selection scheme (39.37) is called, we shall find it convenient to start afresh with a more direct approach.

Stratified random sampling

39.15 As in 39.14, we suppose that the population has been divided into k strata, the l th of these containing N_l individuals, $\sum_{l=1}^k N_l = N$. We now specialize the general scheme (39.37) by supposing that independently within each stratum the sample of n_l members, $\sum_{l=1}^k n_l = n$, is selected with equal probabilities without replacement. Such a scheme is called *stratified random sampling* without replacement. Because of the independence of the selections in the different strata, the theory is very straightforward. We now denote a member of the l th stratum by y_{li} , and will always reserve the first suffix l for stratum identification.

Clearly, we have $\pi_i = \frac{n_l}{N_l}$, for all i in the l th stratum. Thus the unbiased estimator (39.20) becomes

$$\hat{\mu} = \frac{1}{N} \sum_{l=1}^k \sum_{i=1}^{n_l} \frac{y_{li}}{n_l/N_l} = \frac{1}{N} \sum_l N_l m_l, \quad (39.38)$$

where m_l is the sample mean in the l th stratum, whose true mean is denoted by μ_l (*). Since the m_l are independent, the sampling variance of $\hat{\mu}$ is

$$V(\hat{\mu}) = \frac{1}{N^2} \sum_l N_l^2 \text{var } m_l = \frac{1}{N^2} \sum_l N_l^2 \frac{\sigma_l^2}{n_l} \left(1 - \frac{n_l}{N_l}\right), \quad (39.39)$$

where we have applied (39.10) separately in each stratum, and σ_l^2 is the l th stratum variance. If we similarly write s_l^2 for the sample variance in the l th stratum, (39.12) gives the unbiased estimator of (39.39)

$$V(\hat{\mu}) = \frac{1}{N^2} \sum_l N_l^2 \frac{s_l^2}{n_l} \left(1 - \frac{n_l}{N_l}\right). \quad (39.40)$$

(*) There is no danger of confusion with the notation for moments μ_r , if it is remembered that l always identifies a stratum.

If we wish to make all the π_i equal, as in the discussion of 39.14 which led us to stratified sampling, we must have n_i/N_i constant for all i ; this is called a *uniform sampling fraction* (USF). However, we need not now specialize our investigation to this extent, since (39.38–40) are valid for any choice of the n_i , so that we may have variable sampling fractions. What we must now do is to ask which choice of the n_i is most efficient; more precisely, if $n = \sum_i n_i$ is fixed, how should the integers n_i be chosen to minimize the variance (39.39)? Our treatment follows that of Armitage (1947), although the main results were first obtained by Tschuprow in the 1920's, and later by Neyman.

Choice of stratum sample sizes

39.16 The sampling variance (39.39) may be written

$$\frac{1}{N^2} \sum_i N_i^2 \frac{\sigma_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) \equiv \frac{(N-n)}{nN^2} \sum_i N_i \sigma_i^2 + \frac{1}{N^2} \sum_i n_i \left(\frac{N_i \sigma_i}{n_i} - \frac{\sum_i N_i \sigma_i}{n} \right)^2 - \frac{1}{nN} \sum_i N_i \left(\sigma_i - \frac{\sum_i N_i \sigma_i}{N} \right)^2, \quad (39.41)$$

as the reader may verify by expanding the three terms on the right of (39.41). Of these three terms, only the second, which is non-negative, depends upon the stratum sample sizes n_i at all. Thus the sampling variance will be minimized for choice of the n_i when this term is zero, i.e. when

$$\frac{N_i \sigma_i}{n_i} - \frac{\sum_i N_i \sigma_i}{n} = 0$$

or

$$\frac{n_i}{N_i} = \frac{\sigma_i}{\sum_i N_i \sigma_i / n}. \quad (39.42)$$

Thus minimum sampling variance is attained when the *sampling fraction* n_i/N_i in each stratum is made proportional to the square root of the population variance in that stratum. We call this the *minimum variance allocation* to strata, and denote the estimator in this case by $\hat{\mu}_{MV}$. The sample sizes determined in this way are usually fractional, and in practice the nearest integers to them would be used.

It follows at once that the minimized sampling variance is (39.41) with the second term on the right omitted, i.e. on simplifying,

$$V(\hat{\mu}_{MV}) = \frac{1}{N^2} \left\{ \frac{1}{n} (\sum_i N_i \sigma_i)^2 - \sum_i N_i \sigma_i^2 \right\}. \quad (39.43)$$

The sampling variance is not much affected by small variations of the n_i from the values defined by (39.42), as Exercise 39.21 shows.

39.17 Suppose, on the other hand, that we made all the π_i equal, as in our original discussion, by putting

$$\frac{n_i}{N_i} = \frac{n}{N}, \quad \text{all } i. \quad (39.44)$$

It follows at once from (39.38) that in this case $\hat{\mu}$ reduces to the mean of the complete sample. We denote it by $\hat{\mu}_{\text{USF}}$.

It is now easily seen that the second and third terms on the right of (39.41) are equal in value but of opposite sign, and so cancel. The sampling variance resulting from the USF allocation (39.44) is therefore given by the first term alone on the right of (39.41). This will be greater than the minimized value unless the last term on the right of (39.41) is zero, i.e. unless

$$\sigma_l = \sum_i N_i \sigma_i / N, \quad \text{all } l,$$

which requires that all stratum variances be equal. In this case, of course, (39.44) and (39.42) agree.

39.18 We now compare the sampling variance under a USF allocation to strata, as in 39.17, with that under equal-probabilities random sampling with no stratification, given in (39.10), which we now rewrite(*)

$$V(m_R) = \sigma^2 \left(\frac{1}{n} - \frac{1}{N} \right) \equiv \frac{N-n}{nN(N-1)} \left\{ \sum_i (N_i - 1) \sigma_i^2 + \sum_i N_i (\mu_i - \mu)^2 \right\};$$

this identity holds because of the Analysis of Variance identity

$$(N-1) \sigma^2 \equiv \sum_i (N_i - 1) \sigma_i^2 + \sum_i N_i (\mu_i - \mu)^2,$$

which is (35.25) rewritten in another notation. We have seen in 39.17 that for a USF allocation to strata,

$$V(\hat{\mu}_{\text{USF}}) = \frac{N-n}{nN^2} \sum_i N_i \sigma_i^2, \quad (39.45)$$

and this is to be compared with $V(m_R)$. Their difference is

$$V(m_R) - V(\hat{\mu}_{\text{USF}}) = \frac{N-n}{nN} \left[\left\{ \frac{\sum_i (N_i - 1) \sigma_i^2}{N-1} - \frac{\sum_i N_i \sigma_i^2}{N} \right\} + \frac{\sum_i N_i (\mu_i - \mu)^2}{N-1} \right]. \quad (39.46)$$

The term in braces on the right of (39.46) is negative, since it equals

$$\frac{1}{N(N-1)} \sum_i (N_i - N) \sigma_i^2 < 0. \quad (39.47)$$

It therefore follows at once from (39.46) that if the last term on its right is zero, i.e. if all μ_i are equal,

$$V(m_R) < V(\hat{\mu}_{\text{USF}}),$$

so that stratification with USF allocation results in an *increase* in sampling variance. Furthermore, we have already seen in 39.17 that if the σ_i are all equal, the USF is the same as the MV allocation. Thus if the σ_i are all equal and the μ_i all equal,

$$V(m_R) < V(\hat{\mu}_{\text{USF}}) = V(\hat{\mu}_{\text{MV}}). \quad (39.48)$$

In these circumstances, *any* allocation of the sample over strata results in an increase in the sampling variance of the estimator of the population mean.

(*) The suffix *R* denotes equal-probabilities random sampling without replacement.

Even if the μ_i differ slightly and the σ_i differ slightly, a result like (39.48) is still possible, i.e. we may have

$$V(m_R) < V(\hat{\mu}_{MV}) < V(\hat{\mu}_{USF}).$$

Armitage (1947) gives details, some of which are in Exercises 39.14–15.

39.19 Results like (39.48) very rarely occur in sampling practice, for they depend upon the inequality (39.47). Now the left-hand side of (39.47) is of order N^{-1} for fixed k , whereas the other term inside the square brackets in (39.46), $\sum N_i(\mu_i - \mu)^2/(N-1)$, is of order 1 if the N_i are of the same order of magnitude as N . Thus, $N \rightarrow \infty$ with N_i/N fixed, the term in braces on the right of (39.46) is relatively negligible, and the other term is non-negative, so that we have

$$V(m_R) \geq V(\hat{\mu}_{USF}) \geq V(\hat{\mu}_{MV}), \quad (39.49)$$

the equality on the left being attained if and only if all the μ_i are equal, and the equality on the right if and only if all the σ_i are equal.

(39.49) is just what our intuitive argument in 39.14 led us to expect. Both there and more generally in 39.16–18 we have seen that it is the variation among the μ_i which produces the improvement in sampling variance through the use of a stratified sample with a USF (and the variation among the σ_i which produces any further improvement due to MV allocation). This matches with the general conclusion at the end of 39.14, where only a USF was discussed, that strata should be as different between themselves as possible.

The use of strata in sample survey designs has obvious similarity to the use of blocks in experiment design (cf. 38.14). In each case, the grouping (of individuals, of experimental units) has as its aim the elimination from error (sampling, experimental) of the variation between groups (strata, blocks). There is, however, a difference of purpose as well as a similarity of method. In surveys, we are interested primarily in estimating the overall population mean, while the general mean is rarely of interest in experimental situations. This difference is a reflection of the fact that (cf. 38.3) experiments are concerned with hypothetical, rather than existent, populations.

Minimum variance allocation for fixed total cost

39.20 Before leaving the question of sample allocation to strata, we generalize the MV allocation formula (39.42) to take account of variation in costs of sampling between strata. We have deferred this generalization because we may now deduce it as a special case of a general result on minimum variance allocation for fixed total cost, which we shall also find useful in other connexions.

Suppose that in a given sampling problem the sampling variance of the estimator being used is of the form

$$V = v_0 + \sum_{l=1}^k \frac{v_l}{w_l}, \quad (39.50)$$

where v_0 and the v_l are functions of population quantities only and the w_l are not functions of the v 's. Quite commonly the total cost of carrying out the sample survey is representable in the form

$$C = c_0 + \sum_{l=1}^k w_l c_l \quad (39.51)$$

where c_0 is appropriately labelled "overhead cost," and c_l is a cost within an l th category. We now write

$$(V - v_0)(C - c_0) = \sum_l \frac{v_l}{w_l} \sum_l w_l c_l \geq \left\{ \sum_l (v_l c_l) \right\}^2, \quad (39.52)$$

by the Cauchy inequality (see 2.7). The equality in (39.52) is attained if and only if the condition

$$\left(\frac{v_l}{w_l} \right) / (w_l c_l) = \text{constant, all } l,$$

holds. The extreme right-hand side of (39.52) is independent of the w_l , so choice of the w_l to satisfy this condition, which we rewrite

$$w_l^2 \propto v_l / c_l, \text{ all } l, \quad (39.53)$$

will minimize VC , i.e. it will minimize V for fixed C (or C for fixed V).

39.21 In our present application, the variance is given by (39.39), and the cost function is

$$C = c_0 + \sum_{l=1}^k n_l c_l \quad (39.54)$$

where c_0 is the overhead cost and c_l is the cost of an observation in the l th stratum. Identifying (39.39) with (39.50) we see that here

$$v_0 = \frac{1}{N^2} \sum_l N_l \sigma_l^2,$$

$$v_l = \frac{N_l^2 \sigma_l^2}{N^2},$$

$$w_l = n_l,$$

and hence (39.53) gives the MV allocation for fixed C as

$$n_l \propto \frac{N_l \sigma_l}{N c_l^{\frac{1}{2}}}. \quad (39.55)$$

The sampling fraction n_l/N_l is now to be made proportional to the square root of the stratum variance divided by the square root of the stratum cost per observation. (39.42), our previous MV allocation formula ignoring costs, is of course a special case of (39.55) when all c_l are equal.

The formation of strata

39.22 The reader will no doubt have noticed that throughout our discussion from 39.15 onwards, we have been assuming that k fixed strata are given to us in advance, and that our only problem has been how to allocate our sample over these strata. But these strata must have been formed at some stage. How is this best done from the standpoint of ultimately minimizing variance in estimation? We first discuss this with k fixed, and later consider the effect of varying k .

In view of the conclusion of 39.19, the aim in forming strata must be to maximize the variation between the stratum means. Thus, if we knew the distribution of the variable y in the population, we should select $(k-1)$ cutting points within its range

to form k strata. How should these cutting points be chosen? The problem is theoretical, in the sense that we never in practice know the true distribution of y ; however, we may have past values of the variable to guide us, or the values of a variable highly correlated with y may be known, so that the solution to the problem is of practical interest. The basic results are due to Dalenius (1950).

39.23 First, consider the case where a USF is to be used. Ignoring constants n , N , and neglecting the difference between N_l and $N_l - 1$, we rewrite the sampling variance (39.45) as

$$V(\hat{\mu}_{\text{USF}}) \propto V = \sum_{l=1}^k \int_{c_{l-1}}^{c_l} (y - \mu_l)^2 f(y) dy, \quad (39.56)$$

where the distribution of y in the population is represented by $f(y)$, $a \leq y \leq b$, and $a = c_0 < c_1 < c_2 \dots < c_{k-1} < c_k = b$, so that c_1, \dots, c_{k-1} are the cutting points in the range of y which determine the strata boundaries. To minimize (39.56) for choice of the c 's, we put

$$0 = \frac{\partial V}{\partial c_l} = \{(c_l - \mu_l)^2 f(c_l) - (c_l - \mu_{l+1})^2 f(c_l)\}, \quad l = 1, 2, \dots, k-1,$$

so that if $f(c_l) \neq 0$, we have the solution

$$(c_l - \mu_l)^2 = (c_l - \mu_{l+1})^2,$$

and since

$$\mu_l < c_l < \mu_{l+1}$$

this implies

$$c_l = \frac{1}{2}(\mu_l + \mu_{l+1}). \quad (39.57)$$

We therefore choose our cutting points so that they are half-way between the means of the strata they form. Given $f(y)$ and k , this is not difficult to achieve numerically.

39.24 If on the other hand, we are to use the MV allocation sample sizes after the strata are formed, (39.43) is to be minimized by choice of the cutting points. The reader can verify by substituting (39.42) into (39.39) that the second term in braces in (39.43) arises only if the sampling fractions n_l/N_l are not negligible. We neglect these fractions. Ignoring constants n , N , (39.43) is then rewritten

$$V(\hat{\mu}_{\text{MV}}) \propto D^2 = \left[\sum_{l=1}^k \left\{ \int_{c_{l-1}}^{c_l} f(y) dy \cdot \int_{c_{l-1}}^{c_l} (y - \mu_l)^2 f(y) dy \right\}^{\frac{1}{2}} \right]^2. \quad (39.58)$$

We need only minimize D , so that we put

$$0 = \frac{\partial D}{\partial c_l} = \frac{f(c_l) \int_{c_{l-1}}^{c_l} (y - \mu_l)^2 f(y) dy + \int_{c_{l-1}}^{c_l} f(y) dy \cdot (c_l - \mu_l)^2 f(c_l)}{2 \left\{ \int_{c_{l-1}}^{c_l} f(y) dy \cdot \int_{c_{l-1}}^{c_l} (y - \mu_l)^2 f(y) dy \right\}^{\frac{1}{2}}} - \frac{f(c_l) \int_{c_l}^{c_{l+1}} (y - \mu_{l+1})^2 f(y) dy + \int_{c_l}^{c_{l+1}} f(y) dy \cdot (c_l - \mu_{l+1})^2 f(c_l)}{2 \left\{ \int_{c_l}^{c_{l+1}} f(y) dy \cdot \int_{c_l}^{c_{l+1}} (y - \mu_{l+1})^2 f(y) dy \right\}^{\frac{1}{2}}},$$

$$l = 1, 2, \dots, k-1.$$

If $f(c_l) \neq 0$, this reduces, on cancelling a factor N_l throughout, and returning to our original notation, to

$$\frac{\sigma_l^2 + (c_l - \mu_l)^2}{\sigma_l} = \frac{\sigma_{l+1}^2 + (c_l - \mu_{l+1})^2}{\sigma_{l+1}}. \quad (39.59)$$

If σ_l^2 is the same for all strata, (39.59) reduces to (39.57), as it must, but in general these conditions are not so easily satisfied as (39.57), since variances as well as means of strata are involved (as we should expect). We therefore seek approximations to (39.59).

39.25 If k is large, we may assume $f(y)$ to be constant within each stratum, say equal to f_l . Then

$$N_l/N = \int_{c_{l-1}}^{c_l} f(y) dy \doteq f_l \int_{c_{l-1}}^{c_l} dy = f_l(c_l - c_{l-1}),$$

and

$$\sigma_l^2 \doteq \frac{1}{12}(c_l - c_{l-1})^2,$$

the variance of a uniform distribution. Thus the expression to be minimized in (39.43) is, to the same order as in 39.24, proportional to

$$\frac{1}{N} \sum_l N_l \sigma_l \doteq \frac{1}{\sqrt{12}} \sum_l f_l (c_l - c_{l-1})^2 = \frac{1}{\sqrt{12}} \sum_l \{f_l^{1/2} (c_l - c_{l-1})\}^2. \quad (39.60)$$

If we now define the transformation

$$z(y) = \int_a^y \{f(t)\}^{1/2} dt, \quad (39.61)$$

(39.60) may be rewritten

$$\frac{1}{N} \sum_l N_l \sigma_l \doteq \frac{1}{\sqrt{12}} \sum_l \{z(c_l) - z(c_{l-1})\}^2.$$

We therefore require to minimize a sum of squares of quantities $z(c_l) - z(c_{l-1})$, whose sum is fixed at $z(b) - z(a)$. The solution is to make $z(c_l) - z(c_{l-1})$ constant. Thus we apply the transformation (39.61) and determine the cutting points as the roots of the equation

$$z(y) = \frac{r}{k} \{z(b) - z(a)\}, \quad r = 1, 2, \dots, k-1. \quad (39.62)$$

Dalenius and Hodges (1957, 1959), to whom this approximation is due, show how to use it numerically and obtain a closer approximation. An alternative approximation given by Ekman (1959) is derivable by writing, more simply,

$$\sum_l N_l \sigma_l \doteq \frac{1}{\sqrt{12}} \sum_l N_l (c_l - c_{l-1}). \quad (39.63)$$

It follows from our discussion above that $\sum_l \{N_l (c_l - c_{l-1})\}^{1/2}$ is constant and we therefore minimize (39.63) by making $\{N_l (c_l - c_{l-1})\}^{1/2}$ constant, or equivalently

$$N_l (c_l - c_{l-1}) = \text{constant}. \quad (39.64)$$

Cochran (1961) examined numerically the use of approximations (39.62) and (39.64) for small k (equal to 2, 3 or 4) and found that they performed consistently well when

applied to eight representative skew distributions. He also discusses other, less satisfactory, approximations to (39.59).

It should be noted that each of the approximate solutions (39.62) and (39.64) implies that $N_l \sigma_l$ is to be constant over all strata. But if this were so, the MV allocation formula (39.42) shows that we should have n_l constant in every stratum. Thus we can expect a stratification with equal sample size in each stratum to give a variance nearly the minimum possible if the cutting points between strata are chosen to minimize (39.39). The derivation of these is left to the reader as Exercise 39.18; Cochran (1961) found that they differ very little from those obtained by use of (39.62) or (39.64).

S. P. Ghosh (1963b) extends the analysis of the problem of locating stratum boundaries to the case where the strata are based on the values of two (correlated) variables. Random formation of strata is discussed in Exercises 40.4-6.

39.26 If we always sample with a USF or with a MV allocation, and stratum sizes are large, (39.49) assures us that we can never increase variance by subdividing strata to form sub-strata, so that one is led logically to the conclusion that a sample of n observations should be selected with $k = n$ strata, one observation from each; or, if we want to use (39.40) to estimate sampling variance (requiring a minimum of 2 observations in each stratum) with $k = \lceil \frac{1}{2}n \rceil$ strata. When n is small, and we have fairly detailed knowledge of the underlying distribution, there is a good deal to be said for doing this; but otherwise the labour involved is hardly likely to be justified, for there is a good deal of empirical evidence that as k increases, the minimum attainable variance declines more and more slowly, and that very often $k = 2, 3$ or 4 is nearly as good as the best. This is due to the fact that our knowledge of the underlying distribution is usually rather imprecise. Cochran (1963) and Dalenius (1953) give numerical examples. A detailed empirical study of the effect of strata formation and sample allocation on estimator variance was made by Hess *et al.* (1966).

39.27 Finally, we remark that the effect of any stratification upon sampling variance can always be estimated after sampling. To do this, we need only use (39.40) to estimate the variance of $\hat{\mu}$ and compare this with an estimate of the variance of m_R , given in suitable form at the beginning of 39.18. From that formula, we see that the only problem is to estimate

$$\begin{aligned} B &= \sum_l N_l (\mu_l - \mu)^2 \\ &= \sum_l N_l \mu_l^2 - \frac{1}{N} (\sum_l N_l \mu_l)^2 \\ &= \sum_l N_l \mu_l^2 - \frac{1}{N} \{ \sum_l N_l^2 \mu_l^2 + \sum_{l \neq r} \sum N_l N_r \mu_l \mu_r \}. \end{aligned} \quad (39.65)$$

Now because

$$\text{we have} \quad E(m_l^2) - \mu_l^2 \equiv V(m_l) = \frac{\sigma_l^2}{n_l} \left(1 - \frac{n_l}{N_l} \right),$$

$$E \left\{ m_l^2 - \frac{s_l^2}{n_l} \left(1 - \frac{n_l}{N_l} \right) \right\} = \mu_l^2$$

and hence, from (39.65), an unbiased estimator of B is

$$\begin{aligned}\hat{B} &= \sum_l \left(N_l - \frac{N_l^2}{N} \right) \left\{ m_l^2 - \frac{s_l^2}{n_l} \left(1 - \frac{n_l}{N_l} \right) \right\} - \frac{1}{N} \sum_{l \neq r} N_l N_r m_l m_r \\ &= \sum_l N_l m_l^2 - \frac{1}{N} \left(\sum_l N_l m_l \right)^2 - \frac{1}{N} \sum_l (N - N_l)(N - n_l) \frac{s_l^2}{n_l}.\end{aligned}\tag{39.66}$$

The estimator of unstratified random sampling variance is therefore, from 39.18,

$$\hat{V}(m_R) = \frac{N-n}{nN(N-1)} \left\{ \sum_l (N_l - 1) s_l^2 + \hat{B} \right\},\tag{39.67}$$

where \hat{B} is defined by (39.66).

Sample designs: clustering

39.28 We were led to the principle of stratification by our discussion in 39.14 of the effects of varying the probabilities π_{ij} while the π_i were all fixed at n/N ; we saw there that it would be profitable to increase some π_{ij} slightly (and reduce the others to compensate this) as compared with their equal-probabilities sampling values. We now ask whether it may not be worth while to press this further and make some of the π_{ij} as large as possible. From their definitions in 39.6 we see that $\pi_i \geq \pi_{ij} \leq \pi_j$ always, so that if all $\pi_i = n/N$, $\pi_{ij} \leq n/N$.

Suppose, then, that we divide the N individuals in the population into N_1 groups, each containing N_2 individuals (so that $N_1 N_2 = N$), and that for all pairs i, j within any group we put $\pi_{ij} = \pi_i = \pi_j = n/N$. There are $N_2(N_2 - 1)$ pairs within each group, and hence $N_1 N_2(N_2 - 1) = N(N_2 - 1)$ pairs i, j for which π_{ij} is thus increased. From (39.16), all the $N(N - 1)$ π_{ij} in the population must add to $n(n - 1)$, so that the $N(N - 1) - N(N_2 - 1) = N(N - N_2)$ pairs i, j whose π_{ij} has not been increased must be allotted values of π_{ij} adding to $n(n - 1) - N(N_2 - 1) \cdot \frac{n}{N} = n(n - N_2)$. If we make

all these values of π_{ij} equal, each will have value $\frac{n(n - N_2)}{N(N - N_2)}$. Suppose that we chose

n to be a multiple of N_2 , say $n_1 N_2$. Then these $\pi_{ij} = \frac{n_1(n_1 - 1)}{N_1(N_1 - 1)}$. Now we recognize

from 39.6 that this is the value which π_{ij} would have in sampling n_1 units out of N_1 using equal-probabilities random sampling. If we realize that the equality of π_{ij} and π_i within groups implies that each group is either selected as a whole or not at all, we see that our present sample design consists simply of dividing the population into N_1 equal groups (called *clusters*) and selecting n_1 of these with equal probabilities at random.

39.29 A special case of cluster sampling is *systematic sampling*, in which the population is arranged (either physically or by means of a list) in a single sequence of $N = N_1 N_2$ individuals. From among the first N_1 of these, a single individual is selected at random with equal probabilities of selection. If the p th individual is thus selected, the systematic sample consists of the individuals in positions $p, p + N_1, p + 2N_1, \dots, p + (N_2 - 1)N_1$. Thus only N_1 samples, each of size N_2 , are possible and there is a

complete formal identity between systematic sampling and the cluster sampling described in 39.28, but a few special features serve to differentiate them in the literature.

First, the fact that systematic sampling employs clusters which are not physically contiguous strongly contrasts it with other forms of clustering which (as the name implies) use sets of "neighbouring" individuals in the physical population. Second, systematic sampling from lists which are effectively in random order is often used as a practically easier substitute for equal-probabilities random sampling; the method is in these circumstances sometimes known as *quasi-random sampling*. But the most important differentiating factor is that in systematic sampling we commonly find that only one of the N_1 possible samples is selected, i.e. $n_1 = 1$. This immediately renders the estimation of sampling variance impossible without the availability of supplementary information of some kind. It seems simpler to insist that $n_1 \geq 2$ so that valid estimation of sampling variance is possible.

Discussions and bibliographies of systematic sampling are given by Cochran (1963) and Yates (1960), both of whom have contributed notably to its theory.

39.30 What effect will cluster sampling have on $V(\hat{\mu})$ at (39.23)? The contributions from the pairs i, j within the same clusters will now actually be negative, since

$\pi_i \pi_j - \pi_{ij} = -\frac{n}{N} \left(1 - \frac{n}{N}\right)$ for these pairs. The contributions from the other pairs, in different clusters, will be positive, since for them $\pi_i \pi_j - \pi_{ij} = \frac{n_1(N_1 - n_1)}{N_1^2(N_1 - 1)}$. The

argument of 39.14 now applies: if we put the large values of $|y_i - y_j|$ in the same cluster with π_{ij} maximized, we should expect (39.23) to be reduced, and perhaps more dramatically than in 39.14, since the larger values of π_{ij} now operate to reduce (39.23), instead of merely contributing nothing to it as in 39.14.

We thus arrive at a general principle of cluster-construction which operates in exactly the opposite direction from the principle of stratification we have discussed in 39.14: form the population into internally *heterogeneous* groups to reduce cluster sampling variance below equal-probabilities random sampling variance.

As with stratification, we shall now abandon the general framework and enter into a particularized discussion of the details, but we make two general points here.

The primary distinction between stratification and clustering is that every stratum is sampled, while clusters themselves are subject to a selection procedure; it is this fact which leads the principles for the two methods in opposite directions. In stratified sampling, the sampling variability is confined within strata and we construct strata to minimize within-strata variability; in cluster sampling, there is only between-cluster variability, since every cluster is sampled entire, and we construct clusters to minimize this.

Secondly, it is worth while emphasizing here, although it is explicit in 39.14 and 39.28, that neither stratified sampling with a USF, nor cluster sampling with all clusters of equal size, make any change in the overall selection probabilities π_i , which are n/N in each case, just as for equal-probabilities random sampling; these methods operate purely by modifying the joint selection probabilities π_{ij} . Of course, in stratified

sampling with variable sampling fractions, the π_i themselves are changed; and the same applies to cluster sampling when clusters are of unequal size. We shall discuss this more general situation below, and at the same time we find it convenient to generalize our discussion in another direction.

Multi-stage sampling

39.31 Cluster sampling presupposes a grouping of the population members, some of the groups then being selected. It is natural to consider the more general situation where the groups (clusters) are the subject of further sampling at a later stage. Retaining the notation of 39.28 as far as possible, we now formulate this more general situation.

The population of N members is grouped into N_1 first-stage units (previously called clusters). The i th such unit contains N_{i2} second-stage units, the j th of which contains N_{ij3} third-stage units. This hierarchical process can be continued indefinitely, but we shall not consider more than three stages, and indeed it is sometimes enough for our purposes to consider only two stages.

With three stages, we have

$$N = \sum_{i=1}^{N_1} \sum_{j=1}^{N_{i2}} N_{ij3}.$$

At the first stage, n_1 (out of N_1) first-stage units are selected, by a method as yet unspecified. Within the i th of these, n_{i2} second-stage units are selected, from the j th of which n_{ij3} third-stage units are selected, and sample size is

$$n = \sum_{i=1}^{n_1} \sum_{j=1}^{n_{i2}} n_{ij3}.$$

We assume that sampling at any stage may be with unequal probabilities; that selection at any stage is independent of selection at other stages, and that sampling within any unit at a given stage is independent of the sampling within other units at that stage.

39.32 We first have to determine the form of the unbiased estimator. The general theory of selection with unequal probabilities in 39.6-8 applies here, and in particular, (39.20) gives an unbiased estimator $\hat{\mu}$, while (39.23) and (39.24) remain valid expressions for the sampling variance of $\hat{\mu}$, and for an estimator of it. The π_i and π_{ij} in these expressions must now refer, of course, to overall probabilities of selection, taking account of all stages of selection. We now relabel the values y_i in the population as y_{ijk} , each suffix corresponding to a division into the sampling units at a stage, so that, for example, y_{846} is a population value in the 8th first-stage unit, in the 4th second-stage unit within that first-stage unit and in the 6th third-stage unit within that again. In this notation, the unbiased estimator (39.20) becomes

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{n_1} \sum_{j=1}^{n_{i2}} \sum_{k=1}^{n_{ij3}} \frac{y_{ijk}}{\pi_{(ijk)}}, \quad (39.68)$$

where $\pi_{(ijk)}$ is now(*) the overall probability of selection of the single value y_{ijk} .

(*) The parentheses in the suffix are to prevent confusion with the joint probabilities π_{ij} previously used.

For example, suppose that sampling is with equal probabilities at every stage. Then

$$\pi_{(ijk)} = \frac{n_1}{N_1} \cdot \frac{n_{i2}}{N_{i2}} \cdot \frac{n_{ij3}}{N_{ij3}}, \quad (39.69)$$

and (39.68) reduces to

$$\hat{\mu} = \frac{N_1}{Nn_1} \sum_{i=1}^{n_1} \frac{N_{i2}}{n_{i2}} \sum_{j=1}^{n_{i2}} \frac{N_{ij3}}{n_{ij3}} \sum_{k=1}^{n_{ij3}} y_{ijk}.$$

If, further, $N_{i2} = N_2$, all i , and $N_{ij3} = N_3$, all i, j , while $n_{i2} = n_2$ and $n_{ij3} = n_3$ similarly, (39.69) reduces to

$$\hat{\mu} = \frac{N_1 N_2 N_3}{N n_1 n_2 n_3} \sum_i \sum_j \sum_k y_{ijk} = m, \quad (39.70)$$

the overall sample mean, since in this case $N = N_1 N_2 N_3$, $n = n_1 n_2 n_3$. (39.70) is intuitively obvious from the symmetry of this situation.

39.33 Similarly, if there are only two stages of sampling, we drop the suffix k and its summation in (39.68), and the equal-probabilities estimator is (39.69) with $(N_{ij3}/n_{ij3}) \sum_k y_{ijk}$ replaced by y_{ij} , reducing to m in the symmetrical case as at (39.70). We can formally derive many two-stage from three-stage formulae by putting $N_{ij3} = n_{ij3} = 1$, or putting $N_1 = n_1 = 1$. This has the effect of making one suffix (and hence its summation) redundant.

39.34 Just as in our treatment of stratified sampling, we shall find it more convenient to make a direct approach in discussing the sampling variance of our estimator in multi-stage sampling, rather than to persist with the general unequal-probabilities notation—this will avoid, for example, the use of symbols like $\pi_{(ijk)(tuv)}$ for the joint probability of selecting two values y_{ijk} , y_{tuv} . We shall consider in detail two special cases, the first of which is that of sampling with equal probabilities at every stage.

39.35 Consider, then, the sampling variance of the estimator $\hat{\mu}$ at (39.69). It is obvious that each stage of sampling contributes to the variability of $\hat{\mu}$. Since sampling is independent at the different stages, we divide the variation of $\hat{\mu}$ into a sequence of conditional variations. First, we consider its variance at the last stage, conditional upon earlier-stage selections being fixed; then we allow the penultimate-stage selections to vary, and so on until the first-stage selections are varied. With three stages, for example, we take the variance at the third stage, conditional upon the first two stages' selections being fixed, then we allow the second-stage selections to vary, and finally allow the first-stage selections to vary. Symbolically, we write this process

$$V(\hat{\mu}) = E(\hat{\mu} - \mu)^2 = E_1[E_2\{E_3(\hat{\mu} - \mu)^2\}]. \quad (39.71)$$

To evaluate (39.71), we shall make use of a general result concerning the variance of a random variable, special cases of which we have already used in earlier chapters. We now prove the result in general.

39.36 Consider a random variable x , and let c be a condition upon the distribution of x . By the multiplication theorem of probability, $P(x) = P_1(x|c)P_2(c)$, where P , P_1 and P_2 are the probabilities of their arguments. Thus

$$E(x) = E_c \{E(x|c)\}, \quad (39.72)$$

expressing symbolically the fact that in finding the expectation of x we may first impose any condition, find the expectation of x given that condition, and then remove the effect of the condition by taking the expectation of the conditional expectation itself. The variance of x is obtained by considering the identity

$$\begin{aligned} E[E(x^2|c) - \{E(x|c)\}^2] + E[\{E(x|c)\}^2 - (E\{E(x|c)\})^2] \\ \equiv E_c [E(x^2|c) - (E\{E(x|c)\})^2] \\ = E(x^2) - \{E(x)\}^2, \end{aligned}$$

by (39.72). By definition, the first term on the left-hand side is the expectation of the conditional variance of x given c ; the second term on the left is the variance of the conditional expectation of x given c ; and the extreme right-hand side is the unconditional variance of x . Thus, symbolically, we have

$$V(x) = E_c \{V(x|c)\} + V\{E(x|c)\}; \quad (39.73)$$

the unconditional variance is the mean of the conditional variance plus the variance of the conditional mean. Note that if $E(x|c)$ does not depend upon c , the second term in (39.73) is zero, and $V(x)$ is simply the mean of the conditional variance.

The result is quite general: for example, it was, in effect, used in 17.35 to establish the Rao-Blackwell method of improving estimators through sufficient statistics.

39.37 Using (39.73), we now see that (39.71) may be written

$$V(\hat{\mu}) = E_{12} \{V(\hat{\mu})\} + V_{12} \{E(\hat{\mu})\}, \quad (39.74)$$

where we write the symbol "12" for the first- and second-stage conditioning. (39.74) displays the required variance in two parts, which we now evaluate separately.

Consider first the value of $V(\hat{\mu})$. At the third (more generally, the last) stage of

selection, each of the $\sum_{i=1}^{n_1} n_{i2}$ selected second-stage units is sampled; the sampling is with equal probabilities, n_{ij3} individuals being selected out of N_{ij3} . This is in effect a stratified sample with each second-stage unit playing the role of a stratum. Defining

$$m_{ij} = \frac{1}{n_{ij3}} \sum_{k=1}^{n_{ij3}} y_{ijk},$$

we know that

$$V(m_{ij}) = \frac{\sigma_{ij}^2}{n_{ij3}} \left(1 - \frac{n_{ij3}}{N_{ij3}}\right),$$

where σ_{ij}^2 is the population variance in the second-stage unit from which this sample of n_{ij3} was drawn. It follows from (39.69) that

$$V_3(\hat{\mu}) = \left(\frac{N_1}{Nn_1}\right)^2 \sum_{i=1}^{n_1} \left(\frac{N_{i2}}{n_{i2}}\right)^2 \sum_{j=1}^{n_{i2}} N_{ij3}^2 \frac{\sigma_{ij}^2}{n_{ij3}} \left(1 - \frac{n_{ij3}}{N_{ij3}}\right). \quad (39.75)$$

We now have to find the expectation of (39.75) when the selections at earlier stages are allowed to vary. In allowing second-stage selections to vary, we obtain

$$EV(\hat{\mu}) = \left(\frac{N_1}{Nn_1}\right)^2 \sum_{i=1}^{n_1} \frac{N_{i2}^2}{n_{i2}^2} \cdot E \left[\frac{1}{n_{i2}} \sum_{j=1}^{n_{i2}} N_{ij3}^2 \frac{\sigma_{ij}^2}{n_{ij3}} \left(1 - \frac{n_{ij3}}{N_{ij3}}\right) \right],$$

and since the expression in square brackets is a sample mean whose expectation is the corresponding population mean, this is

$$= \left(\frac{N_1}{Nn_1}\right)^2 \sum_{i=1}^{n_1} \frac{N_{i2}^2}{n_{i2}^2} \cdot \frac{1}{N_{i2}} \sum_{j=1}^{N_{i2}} N_{ij3}^2 \frac{\sigma_{ij}^2}{n_{ij3}} \left(1 - \frac{n_{ij3}}{N_{ij3}}\right).$$

Similarly, with the first stage varying,

$$\begin{aligned} EEV(\hat{\mu}) &= \frac{N_1^2}{N^2 n_1} \cdot E \left[\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{N_{i2}}{n_{i2}} \sum_{j=1}^{N_{i2}} N_{ij3}^2 \frac{\sigma_{ij}^2}{n_{ij3}} \left(1 - \frac{n_{ij3}}{N_{ij3}}\right) \right] \\ &= \frac{N_1}{N_2 n_1} \sum_{i=1}^{N_1} \frac{N_{i2}}{n_{i2}} \sum_{j=1}^{N_{i2}} N_{ij3}^2 \frac{\sigma_{ij}^2}{n_{ij3}} \left(1 - \frac{n_{ij3}}{N_{ij3}}\right). \end{aligned} \quad (39.76)$$

We have thus evaluated the first term on the right of (39.74).

39.38 For the second term in (39.74), we first need the value of $E(\hat{\mu})$. From (39.69), this is

$$\begin{aligned} E(\hat{\mu}) &= \frac{N_1}{Nn_1} \sum_{i=1}^{n_1} \frac{N_{i2}}{n_{i2}} \sum_{j=1}^{n_{i2}} N_{ij3} E(m_{ij}) \\ &= \frac{N_1}{Nn_1} \sum_{i=1}^{n_1} \frac{N_{i2}}{n_{i2}} \sum_{j=1}^{n_{i2}} N_{ij3} \mu_{ij}, \end{aligned} \quad (39.77)$$

where μ_{ij} is the population mean corresponding to the sample mean m_{ij} . We write $N_{ij3} \mu_{ij} = T_{ij}$ for the total of the y -values in this (i, j) th unit. We now re-apply (39.73) to the second term on the right of (39.74), which then becomes

$$V(\hat{\mu}) = EE\{V(\hat{\mu})\} + E[V\{E(\hat{\mu})\}] + V[E\{E(\hat{\mu})\}]. \quad (39.78)$$

The last term on the right of (39.78) can now be obtained from (39.77), for, as before,

$$\begin{aligned} E\{E(\hat{\mu})\} &= \frac{N_1}{Nn_1} \sum_{i=1}^{n_1} N_{i2} E \left[\frac{1}{n_{i2}} \sum_{j=1}^{n_{i2}} T_{ij} \right] \\ &= \frac{N_1}{Nn_1} \sum_{i=1}^{n_1} \sum_{j=1}^{N_{i2}} T_{ij}, \end{aligned}$$

and hence

$$V[E\{E(\hat{\mu})\}] = \left(\frac{N_1}{N}\right)^2 V \left[\frac{1}{n_1} \sum_{i=1}^{n_1} \left(\sum_{j=1}^{N_{i2}} T_{ij} \right) \right].$$

The variance required is that of a sample mean, and is therefore

$$= \left(\frac{N_1}{N}\right)^2 \frac{\sigma_{T_1}^2}{n_1} \left(1 - \frac{n_1}{N_1}\right), \quad (39.79)$$

where $\sigma_{T_i}^2$ is the variance between the totals of the y -values in the first-stage units, i.e.

$$\sigma_{T_i}^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} \left(T_i - \frac{\sum_{i=1}^{N_1} T_i}{N_1} \right)^2,$$

$$T_i = \sum_{j=1}^{N_{i2}} T_{ij}.$$

where

Finally, we require the middle term on the right of (39.78). From (39.77),

$$\begin{aligned} V \{E(\hat{\mu})\} &= \left(\frac{N_1}{N n_1} \right)^2 \sum_{i=1}^{n_1} N_{i2}^2 V \left[\frac{1}{n_{i2}} \sum_{j=1}^{n_{i2}} T_{ij} \right] \\ &= \left(\frac{N_1}{N n_1} \right)^2 \sum_{i=1}^{n_1} N_{i2}^2 \frac{\sigma_{T_{ij}}^2}{n_{i2}} \left(1 - \frac{n_{i2}}{N_{i2}} \right), \end{aligned} \quad (39.80)$$

where $\sigma_{T_{ij}}^2$ is the variance between second-stage unit totals within the i th first-stage unit, i.e.

$$\sigma_{T_{ij}}^2 = \frac{1}{N_{i2} - 1} \sum_{j=1}^{N_{i2}} \left(T_{ij} - \frac{\sum_{j=1}^{N_{i2}} T_{ij}}{N_{i2}} \right)^2.$$

(39.80) now gives

$$\begin{aligned} E[V \{E(\hat{\mu})\}] &= \frac{N_1^2}{N^2 n_1} E \left[\frac{1}{n_1} \sum_{i=1}^{n_1} N_{i2}^2 \frac{\sigma_{T_{ij}}^2}{n_{i2}} \left(1 - \frac{n_{i2}}{N_{i2}} \right) \right] \\ &= \frac{N_1}{N^2 n_1} \sum_{i=1}^{N_1} N_{i2}^2 \frac{\sigma_{T_{ij}}^2}{n_{i2}} \left(1 - \frac{n_{i2}}{N_{i2}} \right). \end{aligned} \quad (39.81)$$

39.39 Thus, substituting (39.76), (39.79) and (39.81) into (39.78), we finally have

$$\begin{aligned} V(\hat{\mu}) &= \left(\frac{N_1}{N} \right)^2 \frac{\sigma_{T_i}^2}{n_1} \left(1 - \frac{n_1}{N_1} \right) + \frac{N_1}{N^2 n_1} \sum_{i=1}^{N_1} N_{i2}^2 \frac{\sigma_{T_{ij}}^2}{n_{i2}} \left(1 - \frac{n_{i2}}{N_{i2}} \right) \\ &\quad + \frac{N_1}{N^2 n_1} \sum_{i=1}^{N_1} \frac{N_{i2}}{n_{i2}} \sum_{j=1}^{N_{i2}} N_{ij3}^2 \frac{\sigma_{ij}^2}{n_{ij3}} \left(1 - \frac{n_{ij3}}{N_{ij3}} \right). \end{aligned} \quad (39.82)$$

(39.82) shows, as was evident from (39.78) and indeed from intuitive considerations, that there is a contribution to the variance of the estimator from each stage of sampling.

As a special case of (39.82), consider the symmetrical situation

$$\begin{aligned} N_{i2} &= N_2, \text{ all } i; & N_{ij3} &= N_3, \text{ all } i, j; & N &= N_1 N_2 N_3; \\ n_{i2} &= n_2, \text{ all } i; & n_{ij3} &= n_3, \text{ all } i, j; & n &= n_1 n_2 n_3. \end{aligned}$$

The estimator $\hat{\mu}$ here reduced at (39.70) to the overall sample mean. (39.82) reduces to

$$V(\hat{\mu}) = \frac{\left(1 - \frac{n_1}{N_1} \right) \sigma_{T_i}^2}{(N_2 N_3)^2 n_1} + \frac{\left(1 - \frac{n_2}{N_2} \right)}{N_1 N_3^2 n_1 n_2} \sum_{i=1}^{N_1} \sigma_{T_{ij}}^2 + \frac{\left(1 - \frac{n_3}{N_3} \right)}{N_1 N_2 n_1 n_2 n_3} \sum_{i=1}^{N_1} \sum_{j=1}^{N_{i2}} \sigma_{ij}^2. \quad (39.83)$$

If we now define

$$\begin{aligned}\mu_i &= T_i/(N_2 N_3), \\ \sigma_1^2 &= \frac{1}{N_1-1} \sum_{i=1}^{N_1} (\mu_i - \mu)^2 \equiv \sigma_{T_1}^2/(N_2 N_3)^2 \\ \sigma_2^2 &= \frac{1}{N_1(N_2-1)} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} (\mu_{ij} - \mu)^2 \equiv \frac{1}{N_1 N_3^2} \sum_{i=1}^{N_1} \sigma_{T_{ij}}^2\end{aligned}$$

and

$$\sigma_3^2 = \frac{1}{N_1 N_2 (N_3-1)} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{k=1}^{N_3} (y_{ijk} - \mu_{ij})^2 \equiv \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sigma_{ij}^2,$$

we may rewrite (39.83) in this symmetrical case as

$$V(\hat{\mu}) = \frac{\sigma_1^2}{n_1} \left(1 - \frac{n_1}{N_1}\right) + \frac{\sigma_2^2}{n_1 n_2} \left(1 - \frac{n_2}{N_2}\right) + \frac{\sigma_3^2}{n_1 n_2 n_3} \left(1 - \frac{n_3}{N_3}\right). \quad (39.84)$$

(39.84) makes the extension to further stages of sampling obvious in the symmetrical case. The general formula for $p \geq 2$ stages is

$$V(\hat{\mu}) = \sum_{r=1}^p \frac{\sigma_r^2}{n_1 n_2 \dots n_r} \left(1 - \frac{n_r}{N_r}\right), \quad (39.85)$$

where

$$\sigma_r^2 = \frac{1}{N_1 N_2 \dots N_{r-1} (N_r-1)} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \dots \sum_{u=1}^{N_r} (y_{ij\dots u} - \mu_{ij\dots t})^2.$$

39.40 If any sampling fraction n_r/N_r is unity, the corresponding term in (39.82) disappears. In particular, if every $n_{ij3} = N_{ij3} = 1$, the last summation on the right vanishes, and the remaining two terms give $V(\hat{\mu})$ for two-stage sampling. If every $n_{i2} = N_{i2}$ in addition, only the first term on the right survives, and we are back at the most general form of cluster sampling with unequal-size clusters. Similarly, the first term of (39.84) applies to equal-size cluster sampling.

There is, in fact, no difficulty in seeing how (39.82) would extend for further stages of sampling. A fourth stage would add to the right-hand side the term

$$\frac{N_1}{N^2 n_1} \sum_{i=1}^{N_1} \frac{N_{i2}}{n_{i2}} \sum_{j=1}^{N_{i2}} \frac{N_{ij3}}{n_{ij3}} \sum_{k=1}^{N_{ij3}} N_{ijk4}^2 \frac{\sigma_{ijk4}^2}{n_{ijk4}} \left(1 - \frac{n_{ijk4}}{N_{ijk4}}\right), \quad (39.86)$$

and σ_{ij}^2 in (39.82) would have to be replaced by $\sigma_{T_{ijk}}^2$, defined as an obvious extension of $\sigma_{T_i}^2$ and $\sigma_{T_{ij}}^2$.

However, it is extremely rare in practice for multi-stage sampling to use equal-probabilities sampling throughout. The reason is, quite simply, that the variances in (39.82) are variances between *totals* of the variable y in the different units. When the units vary considerably in size (i.e. contain widely different members of next-stage units) the effect is to make the variance of $\hat{\mu}$ very large. As we have already seen, this point does not arise in the symmetrical case when all units at any stage are of equal size and the variances can be redefined as variances between means. In general, however, we are obliged to seek some other sampling scheme to reduce the sampling variance of $\hat{\mu}$ to acceptable levels. In fact we achieve this by sampling with varying probabilities. We may, of course, use any sets of probabilities whatever at each stage, with $\hat{\mu}$ defined by (39.68), and calculate $V(\hat{\mu})$ from (39.78); but in general

the terms on the right of (39.78) will be more complicated since they will reflect the varying probabilities at every stage.

A completely general expression for $V(\hat{\mu})$ is formally derived at (39.97-8) below for the purpose of estimating sampling variance. Meanwhile, we consider one sample design which is important in practice.

Sampling with probability proportional to size

39.41 Inspection of (39.68) shows that if we make every $\pi_{(ijk)} = n/N$, the estimator $\hat{\mu}$ will reduce to the sample mean m , exactly as in the equal-probabilities symmetrical case at (39.70). To achieve this, we require only that the probability of selecting the i th unit at the first stage be $n_1 A_i^{(1)}/N$; that the probability of selecting the j th unit from the i th first-stage unit be $n_2 A_{ij}^{(2)}/A_i^{(1)}$; and that the probability at the third stage of y_{ijk} being selected be $n_3/A_{ij}^{(2)}$. We then have

$$\pi_{(ijk)} = n_1 \frac{A_i^{(1)}}{N} \cdot n_2 \frac{A_{ij}^{(2)}}{A_i^{(1)}} \cdot \frac{n_3}{A_{ij}^{(2)}} = \frac{n}{N}.$$

It will be seen that within any penultimate-stage unit, final-stage selection is with equal probabilities, but these probabilities in general will differ between penultimate units. The $A_i^{(1)}$, $A_{ij}^{(2)}$ may be any convenient sets of members we choose. A sample design which satisfies $\pi_{(ijk)} = n/N$ is said to be *self-weighting*, since $\hat{\mu}$ is the (equally weighted) mean of the sample.

One simple and convenient choice is to make

$$A_i^{(1)} = \sum_{j=1}^{N_{i2}} N_{ij3}, \quad A_{ij}^{(2)} = N_{ij3}.$$

Each unit at the first stage then has probability of selection proportional to the number of individuals it contains: the same is true at the second stage; and at the final stage, selection is with equal probabilities. We express this by saying that we sample with *probability proportional to size* (p.p.s.) at each of the earlier stages and with equal probabilities at the last stage.

It is easy to see that for any number $p \geq 2$ of stages, overall selection probabilities for every individual will be equal to n/N if we sample with p.p.s. at all but the last stage, where equal probabilities are to be used.

P.p.s. sampling was first theoretically investigated by Hansen and Hurwitz (1943), and was actually the earliest form of explicit unequal-probabilities sampling.

39.42 The simplest way of achieving the self-weighting p.p.s. sample design discussed in 39.41 is to select n_1 first-stage units with replacement, using probabilities

$$p_i^{(1)} = \sum_{j=1}^{N_{i2}} N_{ij3}/N \text{ at each drawing; then to select } n_2 \text{ second-stage units with replacement from each of the } n_1 \text{ selected first-stage units, using probabilities } p_{ij}^{(2)} = N_{ij3}/\sum_{j=1}^{N_{i2}} N_{ij3}$$

at each drawing; and finally to select n_3 third-stage units without replacement from each of the $n_1 n_2$ selected second-stage units with equal probabilities $p_{jik}^{(3)} = 1/N_{ij3}$ at each drawing. The sampling with replacement at the two p.p.s. stages enables us to use the simplified theory of 39.12 in what follows.

We write the estimator

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} y_{ijk} = \frac{1}{n_1 n_2} \sum_i \sum_j m_{ij}.$$

Its variance is given by (39.78), whose terms we now evaluate.

39.43 Because the final-stage sampling is with equal probabilities we have, as in 39.37,

$$V(m_{ij}) = \frac{\sigma_{ij}^2}{n_3} \left(1 - \frac{n_3}{N_{ij3}}\right)$$

so that

$$V(\hat{\mu}) = \frac{1}{(n_1 n_2)^2 n_3} \sum_i \sum_j \sigma_{ij}^2 \left(1 - \frac{n_3}{N_{ij3}}\right).$$

Hence

$$E V(\hat{\mu}) = \frac{1}{n_1^2 n_2 n_3} \sum_{i=1}^{n_1} E \left[\frac{1}{n_2} \sum_{j=1}^{n_2} \sigma_{ij}^2 \left(1 - \frac{n_3}{N_{ij3}}\right) \right]. \quad (39.87)$$

At the second stage, n_2 out of N_{i2} units are selected within the i th first-stage unit, with probabilities $N_{ij3} / \sum_{j=1}^{N_{i2}} N_{ij3}$ at each drawing. Thus (39.87) becomes

$$E V(\hat{\mu}) = \frac{1}{n_1^2 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{N_{i2}} \left(\frac{N_{ij3}}{\sum_j N_{ij3}} \right) \sigma_{ij}^2 \left(1 - \frac{n_3}{N_{ij3}}\right).$$

Similarly

$$\begin{aligned} E E V(\hat{\mu}) &= \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{N_1} \left(\frac{\sum_j N_{ij3}}{N} \right) \sum_{j=1}^{N_{i2}} \left(\frac{N_{ij3}}{\sum_j N_{ij3}} \right) \sigma_{ij}^2 \left(1 - \frac{n_3}{N_{ij3}}\right) \\ &= \frac{1}{n_1 n_2 n_3 N} \sum_{i=1}^{N_1} \sum_{j=1}^{N_{i2}} N_{ij3} \sigma_{ij}^2 \left(1 - \frac{n_3}{N_{ij3}}\right). \end{aligned} \quad (39.88)$$

This is the first term on the right of (39.78). Further,

$$\begin{aligned} E(\hat{\mu}) &= \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mu_{ij}, \\ E E(\hat{\mu}) &= \frac{1}{n_1} \sum_{i=1}^{n_1} E \left[\frac{1}{n_2} \sum_{j=1}^{n_2} \mu_{ij} \right] = \frac{1}{n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{N_{i2}} \left(\frac{N_{ij3}}{\sum_j N_{ij3}} \right) \mu_{ij} \\ &= \frac{1}{n_1} \sum_{i=1}^{n_1} \mu_i, \end{aligned} \quad (39.89)$$

where μ_i , as previously, is the mean of y in the i th first-stage unit. Thus

$$\begin{aligned} V E E(\hat{\mu}) &= V \left(\frac{1}{n_1} \sum_{i=1}^{n_1} \mu_i \right) = \frac{1}{n_1} V(\mu_i) \\ &= \frac{1}{n_1} \sum_{i=1}^{N_1} \left(\frac{\sum_{j=1}^{N_{i2}} N_{ij3}}{N} \right) \left(\mu_i - \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_{i2}} N_{ij3} \mu_i}{N} \right)^2 \\ &= \frac{1}{n_1 N} \sum_{i=1}^{N_1} N_{iT} (\mu_i - \mu)^2, \end{aligned} \quad (39.90)$$

where we now write $N_{iT} = \sum_{j=1}^{N_{i2}} N_{ij3}$ for the total number of individuals in the i th first-stage unit. (39.90) is the last term on the right of (39.78). The middle term there is, from (39.89),

$$\begin{aligned} E V E(\hat{\mu})_{123} &= E \left[\sum_{i=1}^{n_1} V \left\{ \frac{1}{n_1 n_2} \sum_{j=1}^{n_2} \mu_{ij} \right\} \right] = E \left[\frac{1}{n_1^2} \sum_{i=1}^{n_1} \frac{1}{n_2} V(\mu_{ij}) \right] \\ &= E \left[\frac{1}{n_1^2} \sum_{i=1}^{n_1} \frac{1}{n_2} \sum_{j=1}^{N_{i2}} \left(\frac{N_{ij3}}{N_{iT}} \right) \left(\mu_{ij} - \sum_{j=1}^{N_{i2}} \frac{N_{ij3}}{N_{iT}} \mu_{ij} \right)^2 \right] \\ &= E \left[\frac{1}{n_1^2 n_2} \sum_{i=1}^{n_1} \frac{1}{N_{iT}} \sum_{j=1}^{N_{i2}} N_{ij3} (\mu_{ij} - \mu_i)^2 \right] \\ &= \frac{1}{n_1 n_2 N} \sum_{i=1}^{N_1} \sum_{j=1}^{N_{i2}} N_{ij3} (\mu_{ij} - \mu_i)^2. \end{aligned} \quad (39.91)$$

Putting (39.88), (39.90) and (39.91) into (39.78), we obtain

$$\begin{aligned} V(\hat{\mu}) &= \frac{1}{n_1 N} \sum_{i=1}^{N_1} N_{iT} (\mu_i - \mu)^2 + \frac{1}{n_1 n_2 N} \sum_{i=1}^{N_1} \sum_{j=1}^{N_{i2}} N_{ij3} (\mu_{ij} - \mu_i)^2 \\ &\quad + \frac{1}{n_1 n_2 n_3 N} \sum_{i=1}^{N_1} \sum_{j=1}^{N_{i2}} N_{ij3} \sigma_{ij}^2 \left(1 - \frac{n_3}{N_{ij3}} \right). \end{aligned} \quad (39.92)$$

39.44 It will be observed that (39.92) is almost exactly of the same form as (39.84), the variance formula for equal-probabilities sampling in the symmetrical case. In fact, apart from the simplification occasioned by the sampling being now with replacement at the first two stages, the only difference is that the N_{ij3} occur as weights in each component of the variance, as they must do because of the unequal sizes of units. We may write (39.92) in the same form as (39.84),

$$V(\hat{\mu}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1 n_2} + \frac{\sigma_3^2}{n_1 n_2 n_3} \quad (39.93)$$

with obvious definitions, and all the variances are between means, not totals as in (39.82). Thus the present p.p.s. sample design has the effect of eliminating the influence of the varying sizes of the units at the stages of sampling before the last.

Clearly, a similar result will follow for any number of stages. The two-stage result is obtained from (39.92) by putting $n_1 = N_1 = 1$ and making appropriate changes in notation. The reader is asked in Exercise 39.24 to show that in this case, on defining symbols obviously,

$$V(\hat{\mu}) = \frac{1}{n_1 N} \sum_{i=1}^{N_1} N_{i2} (\mu_i - \mu)^2 + \frac{1}{n_1 n_2 N} \sum_{i=1}^{N_1} N_{i2} \sigma_i^2 \left(1 - \frac{n_2}{N_{i2}} \right). \quad (39.94)$$

Estimation of sampling variance in multi-stage sampling

39.45 Although, as indicated at the end of 39.40, the general formula for $V(\hat{\mu})$, for completely arbitrary probabilities of selection at each stage, is lengthy and of no particular interest, it is a remarkable fact that a very general method for the unbiased estimation of the sampling variance of an estimator in multi-stage sampling is easily obtained, and that it is of a very simple form.

Suppose that there is an arbitrary number of stages of sampling, and that we sample without replacement at the first stage, where the probability of the i th unit being included among the n_1 units selected is $\pi_i^{(1)}$, while the joint probability of both i th and j th units being selected at the first stage is $\pi_{ij}^{(1)}$. Using (39.73), we write the variance of any estimator (not necessarily of μ) as

$$V(\hat{\theta}) = E \{ V(\hat{\theta}) \}_{>1} + V \{ E(\hat{\theta}) \}_{>1}, \quad (39.95)$$

where we use the omnibus symbol " >1 " to represent all stages of sampling after the first. We suppose that the estimator $\hat{\theta}$ may be written in the form

$$\hat{\theta} = \sum_{i=1}^{n_1} t_i.$$

(39.68) is of this form, and so more generally is (39.20) for any number of stages. If we apply the alternative estimator (39.27) to the first-stage sampling, therefore writing n_1 for n and interpreting $y_{(r)}$ as the sum of y in the r th first-stage unit, this is another estimator of the form we are discussing. So is its improved form in 39.11.

Since later stages of sampling are carried out independently within the different selected first-stage units,

$$V(\hat{\theta})_{>1} = \sum_{i=1}^{n_1} V(t_i)_{>1},$$

and hence (39.95) may be rewritten

$$\begin{aligned} V(\hat{\theta}) &= V \{ E(\hat{\theta}) \}_{>1} + E \left\{ \sum_{i=1}^{n_1} V(t_i)_{>1} \right\} \\ &= V \left\{ \sum_{i=1}^{n_1} E(t_i)_{>1} \right\} + \sum_{i=1}^{N_1} \pi_i^{(1)} V(t_i)_{>1}, \end{aligned} \quad (39.96)$$

using (39.19). Applying (39.18) to the first term on the right of (39.96), and expanding the other term, we obtain

$$\begin{aligned} V(\hat{\theta}) &= \sum_{i=1}^{N_1} \pi_i^{(1)} (1 - \pi_i^{(1)}) \{ E(t_i)_{>1} \}^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{N_1} (\pi_{ij}^{(1)} - \pi_i^{(1)} \pi_j^{(1)}) E(t_i)_{>1} E(t_j)_{>1} \\ &\quad + \sum_{i=1}^{N_1} \pi_i^{(1)} [E(t_i^2)_{>1} - \{ E(t_i)_{>1} \}^2] \end{aligned} \quad (39.97)$$

$$\begin{aligned} &= \sum_{i=1}^{N_1} \pi_i^{(1)} (1 - \pi_i^{(1)}) E(t_i^2)_{>1} + \sum_{\substack{i,j=1 \\ i \neq j}}^{N_1} (\pi_{ij}^{(1)} - \pi_i^{(1)} \pi_j^{(1)}) E(t_i)_{>1} E(t_j)_{>1} \\ &\quad + \sum_{i=1}^{N_1} (\pi_i^{(1)})^2 V(t_i)_{>1}. \end{aligned} \quad (39.98)$$

39.46 We now seek an unbiased estimator of (39.98). So far as the first two terms on its right are concerned, we need only substitute t_i^2 for $E(t_i^2)_{>1}$ and $t_i t_j$ for $E(t_i)_{>1} E(t_j)_{>1}$, since t_i and t_j are independent at stages of sampling after the first. If we do this, the first two terms also become equal to the first two on the right of (39.97)

with $E(t_i)$ replaced by t_i . Since these latter represent $V_1 \left\{ \sum_{i=1}^{n_1} E(t_i) \right\}$ in (39.96), it follows that we have

$$\sum_{i=1}^{N_1} \pi_i^{(1)} (1 - \pi_i^{(1)}) t_i^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{N_1} (\pi_{ij}^{(1)} - \pi_i^{(1)} \pi_j^{(1)}) t_i t_j = V_1 \left\{ \sum_{i=1}^{n_1} t_i \right\} = V_1(\hat{\theta}). \quad (39.99)$$

The last term on the right of (39.98) is also easily estimated, since if

$$\begin{aligned} E \left\{ \hat{V}_{>1}(t_i) \right\} &= V_{>1}(t_i), \\ E \left\{ \sum_{i=1}^{n_1} \pi_i^{(1)} \hat{V}_{>1}(t_i) \right\} &= E_1 \left[\sum_{i=1}^{n_1} \pi_i^{(1)} E \left\{ \hat{V}_{>1}(t_i) \right\} \right] \\ &= E_1 \left[\sum_{i=1}^{n_1} \pi_i^{(1)} V_{>1}(t_i) \right] = \sum_{i=1}^{N_1} (\pi_i^{(1)})^2 V_{>1}(t_i) \end{aligned}$$

by (39.19). Thus, using (39.99), we see that (39.98) is the expected value of

$$V^*(\hat{\theta}) = V_1(\hat{\theta}) + \sum_{i=1}^{n_1} \pi_i^{(1)} \hat{V}_{>1}(t_i). \quad (39.100)$$

However, (39.100) is not a statistic, since we have yet to estimate $V_1(\hat{\theta})$. If $\hat{V}_1(\hat{\theta})$ is unbiased for $V_1(\hat{\theta})$, we finally have the unbiased estimator of (39.98)

$$\hat{V}(\hat{\theta}) = \hat{V}_1(\hat{\theta}) + \sum_{i=1}^{n_1} \pi_i^{(1)} \hat{V}_{>1}(t_i). \quad (39.101)$$

(39.101) expresses the rule first generally formulated by Durbin (1953) after an earlier more specialized statement by Yates. We state it in words:

An unbiased estimator of sampling variance in multi-stage sampling, when the first-stage sampling is without replacement, is obtainable as the sum of two components. The first component estimates the variance as if only the first-stage sampling had taken place. The second component is the weighted sum of the estimates, within the selected first-stage units, of the variances due to later stages of sampling (the first-stage units being regarded as fixed); the weights are the probabilities of selection of these first-stage units.

39.47 The expression (39.101) may be broken down into further components to facilitate its use. If we write $t_i = \sum_{j=1}^{n_{i2}} t_{ij}$, we may apply (39.101) itself to the terms $\hat{V}_{>1}(t_i)$ and obtain

$$\hat{V}_{>1}(t_i) = \hat{V}_2(t_i) + \sum_{j=1}^{n_{i2}} \pi_{ij}^{(2)} \hat{V}_{>2}(t_{ij}), \quad (39.102)$$

where $\pi_{ij}^{(2)}$ is the probability of selecting the j th second-stage unit in the i th first-stage unit. Substituting (39.102) into (39.101), we obtain

$$\hat{V}(\hat{\theta}) = \hat{V}_1(\hat{\theta}) + \sum_{i=1}^{n_1} \pi_i^{(1)} \hat{V}_2(t_i) + \sum_{i=1}^{n_1} \pi_i^{(1)} \sum_{j=1}^{n_{i2}} \pi_{ij}^{(2)} \hat{V}_{>2}(t_{ij}). \quad (39.103)$$

The pattern for further extension is now obvious. For $p \geq 2$ stages, the result is

$$\hat{V}(\hat{\theta}) = \hat{V}_1(\hat{\theta}) + \sum_{r=2}^p \left[\sum_{i=1}^{n_1} \pi_i^{(1)} \sum_{j=1}^{n_{i2}} \pi_{ij}^{(2)} \dots \sum_{l=1}^{n_{ij \dots l-1}} \pi_{ij \dots l}^{(r-1)} \hat{V}_{>r}(t_{ij \dots l}) \right]. \quad (39.104)$$

It follows from (39.104) that if all $\pi_i^{(1)}$ are small enough to be negligible (which can only happen if the number N_1 of first-stage units in the population is large) only the first term on the right-hand side contributes materially to $V(\hat{\theta})$. In this case, the methods used at the stages of sampling after the first do not affect the *form* of the approximate estimator of sampling variance, although they will, of course, affect its *value* since they determine the value of the first term $\hat{V}_1(\hat{\theta})$.

39.48 In the completely symmetrical equal-probabilities case for which $V(\hat{\mu})$ was given at (39.85), its estimator (as the reader is asked to show in Exercise 39.25) is given by (39.104) as

$$\hat{V}(\hat{\mu}) = \frac{s_1^2}{n_1} \left(1 - \frac{n_1}{N_1}\right) + \sum_{r=2}^p \left(\frac{n_1}{N_1} \frac{n_2}{N_2} \cdots \frac{n_{r-1}}{N_{r-1}}\right) \frac{s_r^2}{n_1 n_2 \cdots n_r} \left(1 - \frac{n_r}{N_r}\right), \quad (39.105)$$

where

$$s_r^2 = \frac{1}{n_1 n_2 \cdots n_{r-1} (n_r - 1)} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \cdots \sum_{u=1}^{n_r} (y_{ij, \dots, tu} - m_{ij, \dots, t})^2$$

is the sample correspondent of σ_r^2 defined below (39.85). (39.105) has the same structure as (39.85), save only that every term after the first is multiplied by the product of earlier-stage sampling fractions $\frac{n_1}{N_1} \cdot \frac{n_2}{N_2} \cdots$. Here again, as at the end of 39.47, we see that if n_1/N_1 is negligible,

$$\hat{V}(\hat{\mu}) \doteq \frac{s_1^2}{n_1}, \quad (39.106)$$

irrespective of the methods of sampling used at later stages than the first.

39.49 If we are sampling *with* replacement at the first stage, (39.104) must be modified to take account of the replacement of (39.19) by (39.33). It will be sufficient to reconsider the derivation of (39.101), from which (39.104) followed. We first note that since (39.96) depended upon later stages of sampling being carried out independently in the different selected first-stage units, we must now insist that if a first-stage unit is selected $r > 1$ times, the later stages of sampling must be carried out r times independently within it.

As we have already seen in 39.12, the effect of sampling with replacement on (39.18) is to allow π_{ij} (but not $\pi_i \pi_j$) to have equal suffixes in the double summation. We may therefore absorb the term in π_i^2 from the first summation into the second. Thus, (39.97) must be replaced by

$$\begin{aligned} V(\hat{\theta}) &= \sum_{i=1}^{N_1} \pi_i^{(1)} \{E(t_i)\}_{>1}^2 + \sum_{i,j=1}^{N_1} (\pi_{ij}^{(1)} - \pi_i^{(1)} \pi_j^{(1)}) \{E(t_i)\}_{>1} \{E(t_j)\}_{>1} \\ &\quad + \sum_{i=1}^{N_1} \pi_i^{(1)} [E(t_i^2) - \{E(t_i)\}_{>1}^2] \\ &= \sum_{i=1}^{N_1} \pi_i^{(1)} \{E(t_i^2)\}_{>1} + \sum_{i,j=1}^{N_1} (\pi_{ij}^{(1)} - \pi_i^{(1)} \pi_j^{(1)}) \{E(t_i)\}_{>1} \{E(t_j)\}_{>1}, \end{aligned} \quad (39.107)$$

the analogue of (39.98). As at (39.99), we see that this is the expected value of

$$V(\hat{\theta}) = \sum_{i=1}^{N_1} \pi_i^{(1)} t_i^2 + \sum_{i,j=1}^{N_1} (\pi_{ij}^{(1)} - \pi_i^{(1)} \pi_j^{(1)}) t_i t_j, \quad (39.108)$$

so that here the unbiased estimating statistic is simply

$$\hat{V}(\hat{\theta}) = \hat{V}_1(\hat{\theta}) \quad (39.109)$$

instead of (39.101). No contribution arises from the subsequent stages of sampling, which influence the *value* of $\hat{V}(\hat{\theta})$, but not its *form*. Clearly, this result also replaces the formula (39.104). Thus the Yates-Durbin rule given at the end of 39.46 simplifies if sampling at the first stage is with replacement: only the first component given by the rule should be calculated.

If first-stage sampling is with equal probabilities, (39.109) reduces to (39.106), which now holds exactly. More generally, in view of the remarks in the last paragraph of 39.47 we see that (39.109) may be regarded as the limit of (39.104) when all $\pi_i^{(1)} \rightarrow 0$, just as the estimator of variance in single-stage simple random sampling with replacement may be derived from the without-replacement formula by letting $n/N \rightarrow 0$.

39.50 If the probabilities of selection are the same at each first-stage drawing, the general formula (39.109) can actually be explicitly written down, for the estimator of variance in one-stage unequal-probabilities sampling with replacement has already been given for that case in 39.12. Here, the estimator is $\hat{\theta} = \sum_{i=1}^{n_1} t_i$ instead of (39.34), so that instead of (39.36) we have

$$V(\hat{\theta}) = \hat{V}_1(\hat{\theta}) = \frac{n_1}{n_1 - 1} \sum_{i=1}^{n_1} \left(t_i - \frac{\hat{\theta}}{n_1} \right)^2, \quad (39.110)$$

a remarkably simple form for estimating the sampling variance in multi-stage sampling with any number of stages when the first stage, with replacement, uses the same unequal probabilities at each drawing; the other stages are arbitrary, apart from the independent sampling condition in the first paragraph of 39.49.

Minimum variance allocation in multi-stage sampling

39.51 We first confine ourselves to the situation where, at each stage of sampling, the same number of units is selected from each previous-stage unit. (For three stages this means that $n_{i2} = n_2$, $n_{ij3} = n_3$.) In this case, both the general equal-probabilities formula (39.82) and the p.p.s. result (39.92) are of the form

$$V(\hat{\mu}) = v_0 + \frac{v_1}{n_1} + \frac{v_2}{n_1 n_2} + \frac{v_3}{n_1 n_2 n_3} \quad (39.111)$$

where the v_l are functions of population quantities only. In many applications, a fairly realistic cost function for three-stage sampling is

$$C = c_0 + n_1 c_1 + n_1 n_2 c_2 + n_1 n_2 n_3 c_3, \quad (39.112)$$

where c_0 is overhead cost and c_l is the cost of sampling a single unit at the l th stage.

(39.111-12) are exactly of the form (39.50-1) with $w_l = n_1 \dots n_l$; it follows from (39.53) that we minimize $V(\hat{\mu})$ for fixed C (or C for fixed $V(\hat{\mu})$) by making

$$(n_1 \dots n_l)^2 \propto v_l/c_l, \quad l = 1, 2, 3. \quad (39.113)$$

Taking ratios of (39.113) with successive values of l , we have

$$n_3^2 = \frac{v_3}{c_3} \cdot \frac{c_2}{v_2}, \quad n_2^2 = \frac{v_2}{c_2} \cdot \frac{c_1}{v_1}, \quad (39.114)$$

n_1 is then determined by (39.114) and whichever of (39.111-12) is fixed. This is a notable result, for it implies that later-stage sample sizes are determined by variances and costs *irrespective of total sample size* $n_1 n_2 n_3$, so that if the amount of money available (or the estimation precision desired) in a multi-stage survey changes, only n_1 should be changed. This result clearly holds for any number of stages $p \geq 2$; (39.113) then holds for $l = 1, 2, \dots, p$, and the $(p-1)$ ratios with successive values of l determine n_2, n_3, \dots, n_p as at (39.114), leaving n_1 to be fixed by cost or accuracy considerations as above.

39.52 The result of 39.51 is concerned with the best choice of the (equal) various-stage sample sizes for given probabilities of selection, i.e. the sample design is fixed with only sample sizes at choice. We may now ask a much more difficult question, following Hansen and Hurwitz (1949): which choice of *probabilities of selection* will minimize sampling variance for fixed cost? In the one-stage case, we saw in 39.8 that if probabilities π_i could be made proportional to the values y_i of the variable, sampling variance would be identically zero. The multi-stage situation is more complicated, as the general variance formula (39.98) indicates, for later-stage probabilities come into the reckoning. However, if sampling is with replacement at the first stage, (39.97) is replaced by (39.107). Furthermore (as the reader is asked to show in Exercise 39.27) if the same set of probabilities is used at each first-stage drawing, use of (39.32) reduces (39.107) to

$$\begin{aligned} V(\hat{\theta}) &= \sum_{i=1}^{N_1} \pi_i^{(1)} \left\{ E(t_i) \right\}_{>1}^2 - \frac{1}{n_1} \left\{ \sum_{i=1}^{N_1} \pi_i^{(1)} E(t_i) \right\}_{>1}^2 + \sum_{i=1}^{N_1} \pi_i^{(1)} [E(t_i^2) - \{E(t_i)\}_{>1}^2] \\ &= \sum_{i=1}^{N_1} \pi_i^{(1)} E(t_i^2)_{>1} - \frac{1}{n_1} \left\{ \sum_{i=1}^{N_1} \pi_i^{(1)} E(t_i) \right\}_{>1}^2. \end{aligned} \quad (39.115)$$

This depends only on the $\pi_i^{(1)}$ at the first stage.

39.53 We now restrict ourselves to two-stage sampling, using constant-probabilities drawings with replacement at the first stage (so that (39.115) holds), and equal-probabilities sampling at the second stage, and to a self-weighting design (see 39.41), so that the overall probability of selection for every individual in the population is the same. This is two-stage p.p.s. sampling as in 39.42, and

$$\pi_i^{(1)} \cdot \frac{n_{i2}}{N_{i2}} = \frac{n}{N},$$

so that

$$n_{i2} = \frac{N_{i2}}{\pi_i^{(1)}} \cdot \frac{n}{N}. \quad (39.116)$$

We use a slightly more general cost-function than the two-stage equivalent of (39.112),

$$C = c_0 + n_1 c_1 + \left(\sum_{i=1}^{n_1} N_{i2} \right) c_2 + \left(\sum_{i=1}^{n_1} n_{i2} \right) c'_2, \quad (39.117)$$

which allows two components of cost at the second stage, one proportional to the total size of the first-stage units sampled and the other to the size of sample. ($N_{i2} c_2$ may be regarded as the cost of "preparing" the i th first-stage unit for the next stage of sampling.) If $c_2 = 0$ and $n_{i2} = n_2$, all i , we return to the form (39.112). By (39.116), (39.117) may be written

$$C = c_0 + n_1 c_1 + \sum_{i=1}^{n_1} N_{i2} \left(c_2 + \frac{n}{N} \frac{c'_2}{\pi_i^{(1)}} \right). \quad (39.118)$$

However, (39.118) is a random variable which we cannot fix in advance, so we work instead with its expectation

$$E(C) = c_0 + n_1 c_1 + c_2 \sum_{i=1}^{N_1} \pi_i^{(1)} N_{i2} + n c'_2$$

(using (39.33)) which we rewrite, since $\sum_{i=1}^{N_1} \pi_i^{(1)} = n_1$ by (39.32),

$$E(C) - c_0 - n c'_2 = \sum_{i=1}^{N_1} \pi_i^{(1)} (c_1 + N_{i2} c_2). \quad (39.119)$$

Because the sample is self-weighting, we know from 39.41 that

$$\hat{\mu} = \sum_{i=1}^{n_1} t_i = \frac{1}{n} \sum_{i=1}^{n_1} \sum_{j=1}^{n_{i2}} y_{ij} = m,$$

so that from (39.116)

$$t_i = \frac{1}{n} \sum_{j=1}^{n_{i2}} y_{ij} = \frac{N_{i2}}{N \pi_i^{(1)}} \cdot \frac{1}{n_{i2}} \sum_{j=1}^{n_{i2}} y_{ij} = \frac{N_{i2}}{N} \frac{m_i}{\pi_i^{(1)}},$$

say. Thus (39.115) becomes, in this case,

$$N^2 V(\hat{\mu}) + \frac{1}{n_1} \left\{ \sum_{i=1}^{N_1} E(N_{i2} m_i) \right\}^2 = \sum_{i=1}^{n_1} \frac{1}{\pi_i^{(1)}} E \{ (N_{i2} m_i)^2 \}. \quad (39.120)$$

39.54 We thus see that the expected cost function is linear in the $\pi_i^{(1)}$, while the variance is a linear function of their reciprocals. This was exactly the situation at (39.50-1), so that the argument of 39.20 holds good with the w_i replaced here by the $\pi_i^{(1)}$, the c_i by $(c_1 + N_{i2} c_2)$ and the v_i by $\{E(N_{i2} m_i)^2\}$. It follows at once from (39.53)

that the $\pi_i^{(1)}$ which minimize $V(\hat{\mu})$ for fixed $E(C)$ are given by

$$(\pi_i^{(1)})^2 \propto \frac{\{E(N_{i2} m_i)^2\}}{c_1 + N_{i2} c_2}, \quad i = 1, 2, \dots, N_1. \quad (39.121)$$

The denominator on the right is the total cost of sampling the i th first-stage unit and preparing it for second-stage sampling; the numerator essentially reflects the variability within the i th first-stage unit; the result is therefore of the general form one would expect.

If $n_{i2} = N_{i2}$, $N_{i2}m_i \equiv \sum_{j=1}^{N_{i2}} y_{ij}$, and if $c_2 = 0$, (39.121) then reduces effectively to the one-stage result: $\pi_i^{(1)}$ must be proportional to the total of the y_{ij} in the i th unit. More generally, if $N_{i2}c_2$ is negligible compared to c_1 , and the m_i vary little, we shall have $\pi_i^{(1)} \propto N_{i2}$ approximately, so that we sample with p.p.s. In the contrary case, when c_1 is negligible relative to $N_{i2}c_2$ we see that if the m_i vary little we have $\pi_i^{(1)} \propto N_{i2}^{1/2}$ approximately. Many practical situations lie between these limiting cases.

39.55 The evaluation of the relative efficiencies of multi-stage and one-stage random sampling, and even more the estimation of their efficiencies from a multi-stage sample, is in general much more complicated than the analogous problem for stratified sampling, which was treated in 39.27; Yates (1960) treats a number of special cases. It is extremely rare for a multi-stage sample to decrease sampling variance, and indeed the motive for multi-stage sampling is almost invariably to reduce costs rather than reduce variance directly; the additional resources can, of course, be applied to an increase in sample size. The result of 39.51 makes our point clear; there, one-stage random sampling is seen to be most efficient only if the solution of (39.114) is that n_2 and n_3 be as large as possible, i.e. $n_2 = N_{i2}$, $n_3 = N_{ij3}$ everywhere. Since the N_{i2} and N_{ij3} are usually themselves very large, such a solution requires very large values for the cost and variance ratios in (39.114), which are almost never found in practice.

39.56 Finally, we mention briefly that the benefits of stratification, which we discussed for single-stage sampling in 39.13–27, apply at every stage of a multi-stage sample. Practical multi-stage sample designs therefore frequently incorporate stratification, particularly at the first stage, which often contributes most to the sampling variance. All the foregoing theory applies separately within each stratum, including the Yates–Durbin rule of 39.46 and 39.49 for estimating variance.

EXERCISES

39.1 A simple random sample of n individuals is drawn with replacement from a population with N members, and d distinct population members are included in the sample. Show that the mean of these d distinct values is an unbiased estimator of the population mean, and that its variance is smaller than that of the overall sample mean for $n > 2$, and equal to it for $n = 2$, by proving the inequality

$$E\left(\frac{1}{d}\right) \leq \frac{1}{N} \left(1 + \frac{N-1}{n}\right).$$

(Raj and Khamis, 1958. The same result holds if d is fixed and n a random variable.)

39.2 Two units are selected from a population of N units without replacement, the probability for the i th unit at the first drawing being p_i , $\sum_{i=1}^N p_i = 1$. The probabilities at the second drawing are made proportional to

$$({}_2)p_j = p_j \left(\frac{1}{1-2p_j} + \frac{1}{1-2p_i} \right)$$

if the i th unit was selected at the first drawing.
Show that

$$\sum_{\substack{j=1 \\ j \neq i}}^N ({}_2)p_j = 1 + \sum_{k=1}^N \frac{p_k}{1-2p_k},$$

that $p_i({}_2)p_j$ is symmetric in i and j , and therefore that p_i is also the unconditional probability that the i th unit is selected at the second drawing. Thus in 39.6, $\pi_i = 2p_i$ and

$$\pi_{ij} = 2p_i p_j \left(\frac{1}{1-2p_j} + \frac{1}{1-2p_i} \right) \left(1 + \sum_k \frac{p_k}{1-2p_k} \right)^{-1}.$$

(Durbin, 1965)

39.3 In (39.23), show that if the probabilities π_i are proportional to the values y_i of the variable (which are taken to be positive) $V(\hat{\mu}) = 0$. Show that in this case the estimator (39.24) of the variance is also equal to zero, but that (39.22) becomes

$$\hat{V}_1(\hat{\mu}) = \mu^2 \left\{ 1 - \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^n \frac{\pi_i \pi_j}{\pi_{ij}} - \frac{m}{N\mu} \right\}$$

where m is the sample mean. Hence show that $\hat{V}_1(\hat{\mu})$ can take negative values.

39.4 In (39.24), show that if $\pi_{1j} = \pi_{2j} = \frac{\delta}{N-2}$ for all $j \neq 1, 2$, and $n = 2$ with y_1, y_2 observed,

$$\hat{V}_2(\hat{\mu}) = \frac{(\pi_{12} + \delta)^2 - \pi_{12}}{\pi_{12}(\pi_{12} + \delta)^2} (y_1 - y_2)^2.$$

Hence show that $\hat{V}_2(\hat{\mu})$ can take negative values.

(Durbin, 1953)

39.5 Show that if sampling without replacement with unequal probabilities is carried out so that at the first drawing the i th individual has probability of selection equal to $p_i > 0$, $\sum_i p_i = 1$, while at all subsequent drawings probabilities of selection are equal, we have, in the notation of 39.6,

$$\pi_{ij} = \frac{(n-1)}{(N-1)(N-2)} \{ (N-n)(p_i + p_j) + n - 2 \},$$

$$\pi_i = \frac{1}{N-1} \{ (N-n)p_i + (n-1) \},$$

and hence that the estimator (39.24) is always positive for this selection scheme.
(A. R. Sen, 1953)

39.6 Show that if $N \geq 3$ and $n = 2$ and the sampling is carried out with probabilities p_i at the first drawing as in Exercise 39.5, while the second drawing is carried out with the $(N-1)$ probabilities in the same proportion as at the first drawing, we have

$$\pi_{ij} = p_i p_j \left(\frac{1}{1-p_i} + \frac{1}{1-p_j} \right),$$

$$\pi_i = p_i \left(\frac{1}{1-p_j} + \sum_{\substack{r=1 \\ r \neq i, j}}^N \frac{p_r}{1-p_r} \right),$$

and that the summation in π_i is a minimum for fixed i, j when

$$\frac{p_r}{1-p_r} = \frac{(1-p_i-p_j)}{(N-2)-(1-p_i-p_j)}.$$

Hence show that the estimator (39.24) is always positive for this selection scheme.

(A. R. Sen, 1953)

39.7 For the selection scheme of Exercise 39.5, show that for any set of p_i we must have $\pi_i > \frac{n-1}{N-1}$, all i , and using (39.16) show that only one π_i at most equals unity.

39.8 Show that for equal-probabilities random sampling without replacement both (39.22) and (39.24) reduce to the estimator of variance $\hat{V}(m)$ given in (39.12).

39.9 Show that the z_u defined at (39.25) have variance

$$V(z_u) = \sum_{(1)} p_{(1)} \sum_{(2)} p_{(2)} \dots \sum_{(u-1)} p_{(u-1)} \sum_{(u)} \frac{y_{(u)}^2}{p_{(u)}} \\ - \sum_{(1)} p_{(1)} \sum_{(2)} p_{(2)} \dots \sum_{(u-1)} \left\{ N\mu - \sum_{r=1}^{u-1} y_{(r)}^2 \right\}, \quad u \geq 2,$$

where each summation is over all available units at the indicated drawing. Using this result, show that for $n = 2$, the statistic (39.27) has variance

$$V(\bar{z}) = \frac{1}{4} \left[\left\{ \sum_{(1)} \frac{y_{(1)}^2}{p_{(1)}} - N^2 \mu^2 \right\} + \left\{ \sum_{(1)} p_{(1)} \sum_{(2)} \frac{y_{(2)}^2}{p_{(2)}} - \sum_{(1)} p_{(1)} (N\mu - y_{(1)}^2)^2 \right\} \right].$$

(Raj, 1956)

39.10 Show that if z defined at (39.26) is to reduce to N times the sample mean m when sampling is with equal probabilities, the weights must satisfy

$$\frac{c_u}{c_1} = \frac{N^{(u-1)}}{(N-2)^{(u-1)}}, \quad 2 \leq u \leq n,$$

these $(n-1)$ conditions, together with $\sum c_u = 1$, determining the weights uniquely.

(cf. Raj, 1956)

39.11 To select a sample of n individuals from a population of N individuals y_s , using probabilities π_s ($s = 1, 2, \dots, N$), consider the following procedure:

- (1) Select a value $M \geq$ the largest individual π_s , say π_{\max} , and choose a number r between 0 and M at random with equal probabilities. Select an integer s_1 by the same process from the integers 1 to N . If $r \leq \pi_{s_1}$, accept y_{s_1} for the sample; if $r > \pi_{s_1}$, repeat this entire operation.
- (2) Select further values s_t successively without replacement from the integers 1 to N . In this sequence accept for the sample every y_{s_t} for which the cumulative sum $r + \sum_{u=2}^t \pi_{s_u}$ first exceeds one of the values $M, 2M, 3M, \dots, (n-1)M$.

Show that if $M \leq \frac{1 - \pi_{\max}}{n - 1}$, this procedure selects the sample, with the required π_s , without replacement. Show that operation (1) repeated n times achieves the required π_s with replacement. (Lahiri, 1951; Grundy, 1954)

39.12 x_1, x_2, \dots, x_n are uncorrelated variates with the same mean μ and variances not necessarily equal. Show that $\bar{x} = \sum_{i=1}^n x_i/n$ is an unbiased estimator of μ , and that its sampling variance is unbiasedly estimated by $\sum_i (x_i - \bar{x})^2 / \{n(n-1)\}$.

39.13 Show that the generalizations of binomial and Poisson sampling discussed in 5.10 are special cases of stratified random sampling, and hence derive the variances given at (5.26) and below (5.27).

39.14 In 39.15–18, show that

$$D = V(m_R) - V(\hat{\mu}_{MV}) - \frac{(N-n)}{nN(N-1)} \sum_l N_l (\mu_l - \mu)^2$$

$$= \frac{1}{nN^2(N-1)} (P - Q - R)$$

where

$$P = N\{N \sum N_l \sigma_l^2 - (\sum N_l \sigma_l)^2\} \geq 0,$$

$$Q = n(\sum N_l \sigma_l^2 - N \sum \sigma_l^2) < 0,$$

$$R = N^2 \sum \sigma_l^2 - (\sum N_l \sigma_l)^2 > 0,$$

$P = 0$ holding if and only if all σ_l are equal. Hence, using (39.46), show that as $N \rightarrow \infty$ with N_l/N fixed, the relationship (39.49) holds.

Show further that if $P = 0$ and n is small enough, D is negative, and that if $N-n$ is also small enough,

$$V(m_R) < V(\hat{\mu}_{MV}).$$

(Armitage, 1947)

39.15 In Exercise 39.14, show that if the σ_l are sufficiently unequal, $P-R$ will often be positive, and that then D is a decreasing function of n , so that the reduction in variance through stratification declines as n increases. Hence show that if any n_l in the MV allocation exceeds the corresponding N_l , we should increase the gain from stratification by putting $n_l = N_l$, and distributing only the $n - N_l$ other observations by the MV allocation.

(Armitage, 1947)

39.16 Show that for any sample design in which the sample mean m is unbiased for the population mean μ ,

$$E\left\{\frac{1}{n} \sum_{i=1}^n (y_i - m)^2\right\} = \sigma_y^2 - \text{var } m,$$

where $\sigma_y^2 = E(y - \mu)^2$. Verify the result for random sampling with equal probabilities without replacement.

(The result is due to L. Kish.)

39.17 In 39.24, show that if there are $k = 2$ strata, and the whole of the second stratum is sampled ($n_2 = N_2$), the best choice of a cutting-point c_1 in the range of y to minimize the sampling variance (39.39) is given by

$$c_1 = \mu_1 + \left(\frac{N_1}{n_1} \right)^{\frac{1}{2}} \sigma_1.$$

(cf. Dalenius, 1952)

39.18 Show that if sample size n_l is to be the same in all strata, the cutting-points in the range of y which minimize (39.39) are given (cf. 39.25) by

$$N_l \{ \sigma_l^2 + (c_l - \mu_l)^2 \} = N_{l+1} \{ \sigma_{l+1}^2 + (c_{l+1} - \mu_{l+1})^2 \}.$$

(Cochran, 1961)

39.19 A large random sample is drawn from a population and the individuals classified into k strata *after* selection. Show that if we use the stratified estimator of the population mean given by (39.38) in these circumstances, the expected value of its variance (taking account of the random variation of the stratum sample sizes) is approximately equal to $V(\hat{\mu}_{\text{USF}})$ given at (39.45).

39.20 Show that if stratum sample sizes are chosen to minimize the variance of the estimated sampling variance (39.40), and n_l/N_l is negligible, the MV allocation formula (39.42) is replaced by

$$\frac{n_l}{N_l} = \frac{\sigma_l (\beta_{2l} - 1)^{1/4}}{\sum_l N_l \sigma_l (\beta_{2l} - 1)^{1/4} / n},$$

where β_{2l} is the moment-ratio μ_4/μ_2^2 for the l th stratum. This allocation will therefore differ from (39.42) unless β_{2l} is constant, all l .

(Ross, 1961)

39.21 Show that if stratum sample sizes n_l differ from those defined by the MV allocation formulae (39.42) by amounts Δn_l , $V(\hat{\mu})$ is increased by approximately the factor

$$1 + \frac{1}{n} \sum_l \{ (\Delta n_l)^2 / n_l \}.$$

39.22 Using 39.2, show that in cluster sampling with equal cluster sizes and a single cluster of size n selected, (39.7) gives the sampling variance of the sample mean, ρ being the intra-class correlation coefficient (cf. 26.25-6, Vol. 2) for clusters.

39.23 Show that the generalizations of binomial and Poisson sampling discussed in 5.11 are special cases of two-stage sampling, and hence derive the variances at (5.29) and (5.30).

39.24 Establish (39.94) from (39.92).

39.25 Deduce (39.105) as a special case of (39.104).

39.26 Use Exercise 39.12 to derive the result (39.110) for the estimation of variance in multi-stage sampling where the first stage is sampled with replacement with the same set of unequal probabilities at each drawing.

(cf. Durbin, 1953)

39.27 Using (39.32), show that when the same set of probabilities is used at each first-stage drawing with replacement, the middle term on the right of (39.107) equals

$$-\frac{1}{n_1} \left\{ \sum_{i=1}^{N_1} \pi_i^{(1)} E(t_i) \right\}^2,$$

and hence that the difference between the with-replacement sampling variance (39.107) and the without-replacement variance (39.97) is expressible as

$$D = \frac{n_1 - 1}{n_1} \left\{ \sum_{i=1}^{N_1} \pi_i^{(1)} E(t_i) \right\}^2 - \sum_{\substack{i,j=1 \\ i \neq j}}^{N_1} \pi_{ij}^{(1)} E(t_i) E(t_j),$$

and that this may be positive or negative. If sampling is with equal probabilities at the first stage, show that $D \geq 0$.

Show that if the with-replacement estimator of variance (39.110) is used when sampling is actually carried out without replacement, it has bias exactly equal to $\frac{n_1}{n_1 - 1} D$, so that if without-replacement sampling has the smaller variance ($D > 0$), use of the with-replacement estimator tends to overestimate the variance.

(Durbin, 1953)

39.28 In a multi-stage design, sampling at the s th stage ($s \geq 1$) is with replacement, and sampling at the $(s+1)$ th stage is with equal probabilities. Show (cf. Exercise 39.27) that if any unit selected r times at the s th stage has its $(s+1)$ th-stage sample size multiplied by r , the variance of the estimator (39.68) of the population mean is less than if r independent $(s+1)$ th-stage samples had been selected within the unit.

39.29 In a multi-stage design, sampling at the s th stage ($s \geq 1$) is with replacement. Show that if any unit selected r times at this stage has the $(s+1)$ th stage of sampling carried out only once within it, and a weight of r given to the results, the variance of the estimator (39.68) of the population mean is greater than if r independent $(s+1)$ th-stage samples had been selected within the unit.

39.30 In sampling with unequal probabilities without replacement of n individuals from a population of N , the probability that a given sample of individuals is selected in a particular order is $p_{(s)}$, and the probability that the same sample is selected in any order is $p_s = \sum_{(s)} p_{(s)}$,

the summation being over the $n!$ possible orderings; $\sum_s p_s = 1$, where the summation is over

the $\binom{N}{n}$ possible samples. If $z_{(s)}$ is a statistic which may take account of the order of selection, and $z_s = \sum_{(s)} p_{(s)} z_{(s)} / p_s$, show using (39.72-3) that

$$E(z_s) = E(z_{(s)})$$

and

$$V(z_s) \leq V(z_{(s)}),$$

the last equality holding only when all values of $z_{(s)}$ are the same. Use this result to show that $\bar{z}_{(s)}$ defined at (39.27) can be improved upon as an estimator, and that $\hat{V}(\bar{z}_{(s)})$ defined at (39.31) can similarly be improved upon.

(cf. M. N. Murthy, 1957, and Pathak, 1961a)

39.31 In sampling with unequal probabilities without replacement, suppose that the N population individuals have probabilities of selection at the first drawing equal to ${}_1p_i$, $i = 1$,

$2, \dots, N$, and that at later drawings the probabilities of selection of hitherto undrawn individuals remain in the same proportions as at the first drawing. Show that (39.25) may then be written

$$z_1 = \frac{y_{(1)}}{{}_{(1)}p_{(1)}},$$

$$z_u = \sum_{r=1}^{u-1} y_{(r)} + \frac{y_{(u)}}{{}_{(1)}p_{(u)}} \left(1 - \sum_{r=1}^{u-1} {}_{(1)}p_{(r)} \right), \quad u = 2, 3, \dots, n,$$

and hence that the improved version of $\bar{z}_{(s)}$ given by Exercise 39.30 may be written

$$\bar{z}_s = \sum_{i=1}^n y_i p_{s|i} / p_s$$

where $p_{s|i}$ is the conditional probability of selecting the observed sample given that y_i is selected first.

(M. N. Murthy, 1957)

CHAPTER 40

SAMPLE SURVEY THEORY: SUPPLEMENTARY INFORMATION

40.1 In Chapter 39, we were concerned with problems of sample design. We now turn to a question which arises whatever that design may be, namely the improvement of the efficiency of estimation.

In 39.8 and 39.13 we touched upon the fact that knowledge of a variable highly correlated with that being studied may assist us to choose probabilities of selection, or to construct strata, to make the sampling variance of the estimator small. Such *supplementary information* concerning an auxiliary variable may also be used directly to change the form of the estimator in order to improve its efficiency. (*)

Ratio estimators and their modifications

40.2 Suppose, as in 39.2, that in sampling a finite population with equal probabilities without replacement we wish to estimate the population mean of y , which we now write μ_y , but that we know the value of the population mean of x , μ_x , and can observe x , as well as y , for the sample values. We clearly ought to be able to turn this extra knowledge to good account. We assume $\mu_x \neq 0 \neq \mu_y$.

Two intuitively reasonable estimators of μ_y are

$$\tilde{\mu}_y = \mu_x m_y / m_x \quad (40.1)$$

and

$$\hat{\mu}_y = \mu_x m_{y/x}, \quad (40.2)$$

where m denotes the sample mean of the variable which is its suffix. (40.1) uses the ratio of sample means, and (40.2) the mean of sample ratios, of y and x as a "correction factor" to the known μ_x .

The expectations of (40.1-2) follow at once from observing that by the definition of a covariance C ,

$$\begin{aligned} C\left(\frac{m_y}{m_x}, m_x\right) &= E(m_y) - E\left(\frac{m_y}{m_x}\right)E(m_x) \\ &= \mu_y - \frac{1}{\mu_x} E(\tilde{\mu}_y)\mu_x, \end{aligned}$$

so that

$$E(\tilde{\mu}_y) = \mu_y - C\left(\frac{m_y}{m_x}, m_x\right). \quad (40.3)$$

(*) This is to be distinguished from the use of supplementary information (an instrumental variable), as in 29.33-46, Vol. 2, to achieve identifiability in estimation. It is, however, analogous to the use of a concomitant variable in the Analysis of Covariance (cf. 35.67-8) in so far as the latter reduces residual variation.

Similarly

$$C\left(\frac{y}{x}, x\right) = \mu_y - E\left(\frac{y}{x}\right)\mu_x$$

so that

$$\begin{aligned} E(\hat{\mu}_y) &= \mu_x E\left(\frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}\right) = \mu_x E\left(\frac{y}{x}\right) \\ &= \mu_y - C\left(\frac{y}{x}, x\right). \end{aligned} \quad (40.4)$$

(40.3-4) show that both estimators are in general biased. Furthermore, since a covariance between two variables cannot exceed the product of their standard deviations (by the Cauchy-Schwarz inequality), we see that

$$\begin{aligned} |E(\tilde{\mu}_y) - \mu_y| &\leq \left\{ V\left(\frac{m_y}{m_x}\right) V(m_x) \right\}^{\frac{1}{2}}, \\ |E(\hat{\mu}_y) - \mu_y| &\leq \left\{ V\left(\frac{y}{x}\right) V(x) \right\}^{\frac{1}{2}}. \end{aligned} \quad (40.5)$$

(40.5) shows that there is a radical difference between the estimators, since as sample size $n \rightarrow \infty$, $V\left(\frac{m_y}{m_x}\right)$ and $V(m_x)$, variances of sample means, are of order n^{-1} , and so is the bias in $\tilde{\mu}_y$; no such effect occurs with $\hat{\mu}_y$, since $V\left(\frac{y}{x}\right)$ and $V(x)$ do not depend on n at all. In fact, it is easy to see from their definitions (40.1-2) that $\tilde{\mu}_y$ is a consistent estimator, since $m_y \rightarrow \mu_y$ and $m_x \rightarrow \mu_x$; but that $\hat{\mu}_y \rightarrow \mu_x \mu_{y/x}$, which will not in general be equal to μ_y . The bias in $\tilde{\mu}_y$ is studied in detail in 40.9 below. First, we see how the bias in $\hat{\mu}_y$ may be removed.

40.3 From (40.4) it is clear that we only need an unbiased estimator of $C\left(\frac{y}{x}, x\right)$ to eliminate the bias in $\hat{\mu}_y$. Since y/x is observed for every sample member, we can calculate the sample covariance of y/x and x . By the bivariate analogue of (12.109), it is the k -statistic k_{11} in the sample which is unbiased for K_{11} in the finite population, and thus the unbiased estimator of the covariance in the population is

$$\begin{aligned} \hat{C}\left(\frac{y}{x}, x\right) &= \frac{N-1}{N} \cdot \frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i}{x_i} - m_{y/x}\right)(x_i - m_x) \\ &= \frac{N-1}{N} \cdot \frac{n}{n-1} (m_y - m_{y/x} m_x). \end{aligned} \quad (40.6)$$

Thus, from (40.2) and (40.4), an unbiased estimator of μ_y is

$$\hat{\mu}'_y = \mu_x m_{y/x} + \frac{N-1}{N} \cdot \frac{n}{n-1} (m_y - m_{y/x} m_x), \quad (40.7)$$

first proposed by Hartley and Ross (1954). If $\frac{N-1}{N} \cdot \frac{n}{n-1}$ is negligible, it reduces to

$$\hat{\mu}'_y \doteq m_y - m_{y/x}(m_x - \mu_x). \quad (40.8)$$

40.4 The removal of the bias in $\tilde{\mu}_y$ is less important, since we have seen in 40.2 that it is consistent, but it is worth considering since ratio estimators are sometimes used when n is small. We cannot directly use the device just used in 40.3, for the covariance we now need to estimate is $C\left(\frac{m_y}{m_x}, m_x\right)$ in (40.3), and a single sample supplies only one value of m_y and of m_x . However, a simple approximation is easily obtained.

When $n = 1$, $C\left(\frac{m_y}{m_x}, m_x\right)$ is identical with $C\left(\frac{y}{x}, x\right)$. Moreover, the covariance between sample means of jointly distributed variables is inversely proportional to sample size (this follows, e.g., from Rule 10 for k -statistics in 12.14, or may easily be proved directly), and the same will hold approximately here, where we seek the covariance of one mean and the ratio of another to the first mean. Thus

$$C\left(\frac{m_y}{m_x}, m_x\right) \doteq \frac{1}{n} C\left(\frac{y}{x}, x\right)$$

and using (40.6) we find from (40.3) the approximately unbiased estimator

$$\begin{aligned} \tilde{\mu}'_y &= \tilde{\mu}_y + \frac{1}{n} C\left(\frac{y}{x}, x\right) \\ &= \mu_x \frac{m_y}{m_x} + \frac{N-1}{N} \cdot \frac{1}{n-1} (m_y - m_{y/x} m_x), \end{aligned} \quad (40.9)$$

a result differently obtained by Nieto de Pascual (1961). The absence of the factor n in the second term of (40.9), compared with (40.7), again illustrates the different orders of magnitude of the biases in (40.1-2).

40.5 We now have to examine the variances of the alternative modified ratio estimators (40.7) and (40.9) as a guide to choosing between them in different circumstances.

We consider only the case when $N \rightarrow \infty$, so that sampling is effectively simple random. Using (40.6), we rewrite (40.7) as

$$\hat{\mu}'_y = \mu_x m_{y/x} + k_{11}, \quad (40.10)$$

where k_{11} is the k -statistic of the variables y/x and x . Thus

$$V(\hat{\mu}'_y) = \mu_x^2 V(m_{y/x}) + 2\mu_x C(m_{y/x}, k_{11}) + V(k_{11}), \quad (40.11)$$

which in the notation of 13.2 is written

$$V(\hat{\mu}'_y) = \mu_x^2 V(m_{y/x}) + 2\mu_x \kappa \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} + \kappa \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \quad (40.12)$$

Now by 12.14,

$$\kappa \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \frac{\kappa_{21}}{n} \equiv \frac{\mu_{21}}{n}, \quad (40.13)$$

using (3.80); while (13.7) and (3.81) give

$$\kappa \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \frac{\mu_{22}}{n} + \frac{\mu_{20}\mu_{02}}{n(n-1)} - \frac{(n-2)\mu_{11}^2}{n(n-1)}, \quad (40.14)$$

where the cumulants and moments in (40.13-14) refer to the joint distribution of y/x and x . We therefore rewrite (40.12) in this notation as

$$V(\hat{\mu}'_y) = \mu_{01}^2 \mu_{20}/n + 2\mu'_{01} \mu_{21}/n + \frac{1}{n} \left\{ \mu_{22} + \frac{1}{n-1} (\mu_{20} \mu_{02} - (n-2) \mu_{11}^2) \right\}. \quad (40.15)$$

(40.15) may be usefully simplified. First we observe that by definition

$$\begin{aligned} V \left\{ \left(\frac{y}{x} - \mu_{y/x} \right) (x - \mu_x) \right\} &\equiv E \left\{ \left(\frac{y}{x} - \mu_{y/x} \right)^2 (x - \mu_x)^2 \right\} - \left\{ E \left(\frac{y}{x} - \mu_{y/x} \right) (x - \mu_x) \right\}^2 \\ &= \mu_{22} - \mu_{11}^2 \end{aligned}$$

in the notation of (40.15), which is thus equivalent to

$$nV(\hat{\mu}'_y) = \mu_{01} \mu_{20} + 2\mu'_{01} \mu_{21} + V \left\{ \left(\frac{y}{x} - \mu_{y/x} \right) (x - \mu_x) \right\} + \frac{1}{n-1} (\mu_{20} \mu_{02} + \mu_{11}^2). \quad (40.16)$$

Now consider the identity

$$\mu_x \left(\frac{y}{x} - \mu_{y/x} \right) + \left(\frac{y}{x} - \mu_{y/x} \right) (x - \mu_x) \equiv y - \mu_{y/x} x. \quad (40.17)$$

If we take the variance of the left-hand side of (40.17), we see that it is exactly the first three terms on the right of (40.16). We may therefore replace these by the variance of the right-hand side of (40.17), obtaining

$$nV(\hat{\mu}'_y) = V(y - \mu_{y/x} x) + \frac{1}{n-1} (\mu_{20} \mu_{02} + \mu_{11}^2),$$

and returning to our original notation, this is

$$nV(\hat{\mu}'_y) = V(y) + \mu_{y/x}^2 V(x) - 2\mu_{y/x} C(y, x) + \frac{1}{n-1} \left\{ V \left(\frac{y}{x} \right) V(x) + C^2 \left(\frac{y}{x}, x \right) \right\}, \quad (40.18)$$

a result obtained by Goodman and Hartley (1958). As $n \rightarrow \infty$, the term in braces in (40.18) may be neglected and

$$nV(\hat{\mu}'_y) \sim V(y) + \mu_{y/x}^2 V(x) - 2\mu_{y/x} C(y, x). \quad (40.19)$$

(40.18) is most easily estimated by expressing its form (40.12) in terms of cumulants and using k -statistics to estimate them—Goodman and Hartley (1958) give computing formulae.

Robson (1957) generalized (40.18), and also its unbiased estimator, to take account of the finiteness of the population.

40.6 We may similarly obtain the variance of (40.9), which we rewrite analogously to (40.10) as

$$\tilde{\mu}'_y = \mu_x \frac{m_y}{m_x} + \frac{k_{11}}{n}. \quad (40.20)$$

We see that

$$V(\tilde{\mu}'_y) = \mu_x^2 V \left(\frac{m_y}{m_x} \right) + \frac{2}{n} \mu_x C \left(\frac{m_y}{m_x}, k_{11} \right) + \frac{1}{n^2} V(k_{11}).$$

Just as we did in deriving $\tilde{\mu}'_y$ in 40.4, we approximate by writing

$$C\left(\frac{m_y}{m_x}, k_{11}\right) \doteq \frac{1}{n} C\left(\frac{y}{x}, k_{11}\right),$$

so that

$$\begin{aligned} V(\tilde{\mu}'_y) &\doteq \mu_x^2 V\left(\frac{m_y}{m_x}\right) + \frac{1}{n^2} \left\{ 2\mu_x C\left(\frac{y}{x}, k_{11}\right) + V(k_{11}) \right\} \\ &\doteq \mu_x^2 V\left(\frac{m_y}{m_x}\right) + \frac{1}{n^2} \left\{ V(\hat{\mu}'_y) - \mu_x^2 V(m_{y/x}) \right\} \end{aligned} \quad (40.21)$$

from (40.11). All three variances on the right of (40.21) are of order n^{-1} , so that the second term is of relative order n^{-2} compared with the first term on the right. Our approximation may therefore be written

$$V(\tilde{\mu}'_y) = \mu_x^2 V\left(\frac{m_y}{m_x}\right) \left\{ 1 + O\left(\frac{1}{n}\right) \right\}. \quad (40.22)$$

Since the bias in the unadjusted estimator $\tilde{\mu}_y$ was seen in 40.2 to be of order n^{-1} , the bias in $\tilde{\mu}'_y$ is of no greater order, and its square will be of no greater order than n^{-2} . Thus the mean-square error (more appropriate than the variance in view of the biasedness of $\tilde{\mu}'_y$) may be written

$$E\{(\tilde{\mu}'_y - \mu_y)^2\} = \mu_x^2 V\left(\frac{m_y}{m_x}\right) \left\{ 1 + O\left(\frac{1}{n}\right) \right\}. \quad (40.23)$$

A more precise approximation is given by Nieto de Pascual (1961). The leading term in (40.22) and (40.23) is the variance of the unmodified estimator $\tilde{\mu}_y$, and is easily evaluated to order n^{-1} by using (10.17), which here gives

$$V\left(\frac{m_y}{m_x}\right) = \frac{1}{n} \frac{\mu_y^2}{\mu_x^2} \left\{ \frac{V(y)}{\mu_y^2} + \frac{V(x)}{\mu_x^2} - \frac{2C(y, x)}{\mu_y \mu_x} \right\}. \quad (40.24)$$

Thus (40.23) becomes

$$nE\{(\tilde{\mu}'_y - \mu_y)^2\} \sim V(y) + \frac{\mu_y^2}{\mu_x^2} V(x) - 2\frac{\mu_y}{\mu_x} C(y, x). \quad (40.25)$$

(40.25) may be estimated with slight bias by replacing μ_y/μ_x by m_y/m_x and the variances and covariance by their unbiased estimators. This gives an estimator of the mean-square error

$$\hat{E}\{(\tilde{\mu}'_y - \mu_y)^2\} = \frac{1}{n(n-1)} \sum_{i=1}^n \left(y_i - \frac{m_y}{m_x} x_i \right)^2.$$

40.7 We now compare (40.19) and (40.25). Following Goodman and Hartley (1958), the difference may be written

$$nE\{(\tilde{\mu}'_y - \mu_y)^2\} - nV(\hat{\mu}'_y) = V(x) \left\{ \left(\frac{\mu_y}{\mu_x} - \frac{C(y, x)}{V(x)} \right)^2 - \left(\mu_{y/x} - \frac{C(y, x)}{V(x)} \right)^2 \right\}. \quad (40.26)$$

(40.26) makes it clear that the modified ratio of means estimator $\tilde{\mu}'_y$ is more, or less, efficient than the modified mean of ratios estimator $\hat{\mu}'_y$ according as the linear regression

coefficient of y upon x , $\beta_{yx} = C(y, x)/V(x)$, is nearer to the population ratio of means or to the population mean of ratios. In practice, the former situation seems to be more common, and the slightly biased estimator (40.9) is then preferable to the unbiased (40.7).

(40.9) is more efficient than the ordinary sample mean estimator m_y if the right-hand side of (40.25) is less than its first term, i.e. if

$$\beta_{yx} > \frac{1}{2} \frac{\mu_y}{\mu_x}. \quad (40.27)$$

We have thus characterized the efficiency of $\tilde{\mu}'_y$, compared to both $\hat{\mu}'_y$ and m_y , in terms of the relative magnitudes of the regression coefficient β_{yx} and the population ratio of means μ_y/μ_x . Since $\tilde{\mu}'_y$ essentially estimates by the sample ratio of means m_y/m_x , this is as we should expect.

Olkin (1958) generalizes the theory of the unmodified $\tilde{\mu}_y$ to the case where x is a vector.

40.8 The approximately unbiased estimator (40.9) was obtained by directly estimating the bias in $\tilde{\mu}_y$ given by (40.3). We could, alternatively, have reduced the order of magnitude of the bias by using Quenouille's method, described in 17.10. This would involve the calculation of $\tilde{\mu}_y$ for each of the n different samples of size $(n-1)$ which exclude a single observation, averaging these n values, and using (17.10) to obtain a modified estimator with bias of order n^{-2} and variance unaffected to order n^{-1} (cf. Exercise 17.18), as was seen to be the case for $\tilde{\mu}'_y$ in (40.23).

Durbin (1959a) used a simpler form of Quenouille's method to modify a general type of ratio estimator of form $r = \frac{t_y}{t_x}$ (which includes (40.1) as a particular case), whose bias in estimating $E(t_y)/E(t_x)$ is assumed to be of order n^{-1} . If the same statistic r is calculated for the first $\frac{1}{2}n$ and second $\frac{1}{2}n$ observations (n even) and denoted by r_1, r_2 respectively, the modified estimator is

$$t(r) = 2r - \frac{1}{2}(r_1 + r_2). \quad (40.28)$$

If the regression of t_y on t_x is linear with constant variance of order n^{-1} , and t_x itself is normally distributed with variance of order n^{-1} , (40.28) was shown to have bias of order n^{-2} and variance which agrees to order n^{-1} with that of r but is *smaller* asymptotically when terms of order n^{-2} and lower are taken into account. A similar result holds when t_x has a Gamma distribution.

J. N. K. Rao (1965) shows that use of Quenouille's original method gives even smaller bias and mean-square error.

This result is more general than at first appears, because t_y and t_x will usually be asymptotically bivariate normally distributed with variances of order n^{-1} by the Central Limit theorem—this is certainly true of m_y and m_x in (40.1)—and the linear regression assumption therefore satisfied. It follows that in such a situation there is nothing to be lost by bias-elimination using (40.28).

A simple numerical example with $n = 2$, $N = 4$, discussed by Goodman and Hartley (1958), Durbin (1959a) and Nieto de Pascual (1961), gives the results:

Population values of (x, y) : (2, 2), (2, 6), (4, 6) and (6, 10)

Estimator	Equation	Mean-square error
$t(\mu_y)$	(40.28), (40.1)	
$\tilde{\mu}_y'$	(40.9)	0.38
$\tilde{\mu}_y''$	(40.7)	0.44
$\tilde{\mu}_y$	(40.1)	0.56
μ_y	(40.2)	0.92
m_y	(sample mean)	2.41
		2.67

(40.29)

Exercise 40.17 asks the reader to verify these values.

40.9 If we write (40.1) in the form

$$\frac{\tilde{\mu}_y}{\mu_x} = \frac{m_y}{m_x} = \frac{\mu_y}{\mu_x} \left(1 + \frac{m_y - \mu_y}{\mu_y} \right) \left(1 + \frac{m_x - \mu_x}{\mu_x} \right)^{-1}, \quad (40.30)$$

and expand the negative binomial into a Taylor series, valid with probability 1 as $N, n \rightarrow \infty$, we find on taking expectations

$$E\left(\frac{m_y}{m_x}\right) = \frac{\mu_y}{\mu_x} \left\{ 1 + \left(\frac{1}{n} - \frac{1}{N} \right) \left(\frac{V(x)}{\mu_x^2} - \frac{C(y, x)}{\mu_y \mu_x} \right) + O\left(\frac{1}{M^2}\right) \right\}, \quad (40.31)$$

where M stands for n or N indifferently. Thus the estimator

$$u = \frac{m_y}{m_x} \left\{ 1 - \left(\frac{1}{n} - \frac{1}{N} \right) \left(\frac{s_x^2}{m_x^2} - \frac{s_{xy}}{m_y m_x} \right) \right\} \quad (40.32)$$

has the first-order bias of m_y/m_x removed. In (40.32), the sample variance and covariance are defined with $(n-1)$ as divisor, as usual.

It is a straightforward, though somewhat tedious, matter to evaluate the mean and variance of u , using the results (the first three of which were given at (12.117), (12.119) and (12.121), the remainder being derivable by the methods of Example 13.2):

$$\left. \begin{aligned} E(m_x - \mu_x)^3 &= \left(\alpha_2 - \frac{3\alpha_1}{N} \right) K_{20}, \\ E(m_x - \mu_x)^4 &= 3\alpha_1^2 K_{20}^2 + O(n^{-3}), \\ E(m_x - \mu_x)(s_x^2 - K_{20}) &= \alpha_1 K_{30}, \\ E(m_x - \mu_x)^2(m_y - \mu_y) &= \alpha_2 K_{21}, \\ E(m_x - \mu_x)^3(m_y - \mu_y) &= \alpha_1^2 K_{20} K_{11} + O(n^{-3}), \\ E(s_x^2 - K_{20})(m_y - \mu_y) &= \alpha_1 K_{21}, \\ E(m_x - \mu_x)(s_{xy} - K_{11}) &= \alpha_1 K_{21}, \\ E(s_{xy} - K_{11})(m_y - \mu_y) &= \alpha_1 K_{12}. \end{aligned} \right\} \quad (40.33)$$

Here $\alpha_r = (n^{-r} - N^{-r})$ as at (12.116), and we have dropped the suffix N to E , as throughout this and the last chapter. Tin (1965) gives the results to order n^{-2} ,

$$E(u) = \left(\frac{\mu_y}{\mu_x} \right) \left\{ 1 - \left(2\alpha_2 - \frac{3\alpha_1}{N} \right) (C_{21} - C_{30}) - 3\alpha_1^2 C_{20} (C_{20} - C_{11}) \right\}, \quad (40.34)$$

$$V(u) = \left(\frac{\mu_y}{\mu_x}\right)^2 \left\{ \alpha_1 (C_{20} + C_{02} - 2C_{11}) + \alpha_1^2 (2C_{20}^2 - 4C_{20}C_{11} + C_{11}^2 + C_{20}C_{02}) \right. \\ \left. + \frac{2\alpha_1}{N} (C_{30} - 2C_{21} + C_{12}) \right\}, \quad (40.35)$$

where

$$C_{rs} = K_{rs}/(\mu_x^r \mu_y^s).$$

In a precisely similar way, Tin (1965) gives the results for the simple ratio estimator $\tilde{\mu}_y/\mu_x = m_y/m_x$ as

$$E\left(\frac{m_y}{m_x}\right) = \left(\frac{\mu_y}{\mu_x}\right) \left\{ 1 + \alpha_1 (C_{20} - C_{11}) + \left(\alpha_2 - \frac{3\alpha_1}{N}\right) (C_{21} - C_{30}) + 3\alpha_1^2 C_{20} (C_{20} - C_{11}) \right\}, \quad (40.36)$$

$$V\left(\frac{m_y}{m_x}\right) = \left(\frac{\mu_y}{\mu_x}\right)^2 \left\{ \alpha_1 (C_{20} + C_{02} - 2C_{11}) + \alpha_1^2 (8C_{20}^2 - 16C_{20}C_{11} + 5C_{11}^2 + 3C_{20}C_{02}) \right. \\ \left. - 2\left(\alpha_2 - \frac{3\alpha_1}{N}\right) (C_{30} - 2C_{21} + C_{12}) \right\}, \quad (40.37)$$

while for $t\left(\frac{m_y}{m_x}\right)$, defined by (40.28), he finds

$$E\left\{t\left(\frac{m_y}{m_x}\right)\right\} = \left(\frac{\mu_y}{\mu_x}\right) \left\{ 1 - (C_{20} - C_{11})/N - 2\alpha_2 (C_{21} - C_{30}) \right. \\ \left. + 3\left(\frac{2}{n^2} - \frac{1}{N^2}\right) C_{20} (C_{20} - C_{11}) \right\}, \quad (40.38)$$

$$V\left\{t\left(\frac{m_y}{m_x}\right)\right\} = \left(\frac{\mu_y}{\mu_x}\right)^2 \left\{ \alpha_1 (C_{20} + C_{02} - 2C_{11}) + 2\left(\frac{2}{n^2} - \frac{5}{nN} + \frac{4}{N^2}\right) C_{20} (C_{20} - 2C_{11}) \right. \\ \left. + \left(\frac{2}{n^2} - \frac{6}{nN} + \frac{5}{N^2}\right) C_{11}^2 + \left(\frac{2}{n^2} - \frac{4}{nN} + \frac{3}{N^2}\right) C_{20}C_{02} \right. \\ \left. + \frac{4\alpha_1}{N} (C_{30} - 2C_{21} + C_{12}) \right\}. \quad (40.39)$$

These results make it clear that the bias in u and in $t(m_y/m_x)$ is very small, with no term of order n^{-1} in either, as opposed to the bias in m_y/m_x . All three variances have the same leading term, which we have already encountered at (40.24), where we saw it to be also the mean-square error of $\tilde{\mu}_y'/\mu_x$ defined by (40.9). The reader is left to show in Exercise 40.13 that to the next order of approximation, we have

$$V(u) < V\left\{t\left(\frac{m_y}{m_x}\right)\right\} < V\left(\frac{m_y}{m_x}\right), \quad (40.40)$$

and thus u defined at (40.32) seems preferable to the other estimators considered here on grounds of bias and mean-square error.

Tin (1965) also considers another estimator closely related to u —cf. Exercise 40.14—and makes comparisons in bivariate normal and other situations.

Regression estimators

40.10 Given that we know the population mean μ_x of a supplementary variable, as in 40.2, it is natural to consider the application of the theory of regression to improve

efficiency in the estimation of μ_y . The simplest form is a linear regression estimator

$$\check{\mu}_y = m_y + b(\mu_x - m_x). \quad (40.41)$$

This is a generalization of (40.1), to which it reduces if we choose $b = m_y/m_x$. However, b is usually chosen as the LS regression coefficient of y upon x .

If the linear model (19.8) holds for the relationship between y and x , the LS theory of Chapter 19 and of 28.12 onwards holds; however, we are dealing here with a finite population, and in any case x is a random variable in many applications, so that the LS theory cannot hold exactly. Instead, we observe that m_y and m_x are jointly distributed with means μ_y, μ_x , variances $V(x)/n, V(y)/n$ and covariance $C(y, x)/n$. Thus, from (40.41), if we ignore sampling errors in b , however it is chosen,

$$nV(\check{\mu}_y) = V(y) + b^2V(x) - 2bC(y, x), \quad (40.42)$$

with unbiased estimator

$$\hat{V}(\check{\mu}_y) = \frac{1}{n(n-1)} \sum_{i=1}^n \{(y_i - m_y) - b(x_i - m_x)\}^2. \quad (40.43)$$

The asymptotic formula below (40.25) for the estimated mean-square error of $\tilde{\mu}'_y$, which is also its asymptotic variance and that of the unmodified $\tilde{\mu}_y$, is derivable from (40.43) by putting $b = m_y/m_x$ as above. Under the Central Limit theorem, m_y and m_x will usually be asymptotically normal, and hence $\check{\mu}_y$ is so.

40.11 (40.42) is only an asymptotic result, because only then is the sampling error in b negligible. We see that the regression estimator $\check{\mu}_y$ is more efficient than the sample mean estimator m_y if the right-hand side of (40.42) is less than its first term, i.e. if $2bC(y, x) > b^2V(x)$.

If we choose b as the LS coefficient, which as $n \rightarrow \infty$ tends to $C(y, x)/V(x)$, this condition is always satisfied if $C(y, x) \neq 0$. Similarly, comparing (40.42) with (40.25), we see that the condition for $\check{\mu}_y$ to be more efficient than $\tilde{\mu}'_y$ is

$$V(x) \left\{ b^2 - \left(\frac{\mu_y}{\mu_x} \right)^2 \right\} < 2C(y, x) \left(b - \frac{\mu_y}{\mu_x} \right). \quad (40.44)$$

If b is the LS coefficient and tends to $C(y, x)/V(x)$, (40.44) reduces asymptotically to

$$V(x) \left(b - \frac{\mu_y}{\mu_x} \right)^2 > 0 \quad (40.45)$$

which holds except when $b = \mu_y/\mu_x$ when, as we have seen, (40.41) will reduce to (40.1) asymptotically. There is thus nothing to be lost by using the LS regression estimator, at least asymptotically.

40.12 The regression estimator (40.41) is, of course, biased. To remove this bias, we discuss general methods for constructing unbiased estimators, due to Mickey (1959) and W. H. Williams (1961, 1962), which will also throw light upon our earlier problems in ratio estimation.

Unbiased estimation with a supplementary variable

40.13 We begin from the observation that, for any constant a , the estimator

$$m_y - a(m_x - \mu_x) \quad (40.46)$$

will be unbiased for μ_y , but this will not generally be so if a is a statistic calculated from the same sample as m_x . Suppose now that the sample of n observations is split at random into a subsample of p observations and a "remainder" sample of $(n-p)$ observations. (To be precise, we may choose the first p observations in the order of drawing as the subsample.) We now use the subsample to determine a in (40.46), and calculate the means m_y , m_x for the remainder sample only; since the remainder sample is a random sample from the $(N-p)$ population members not included in the subsample, this will give us an unbiased estimator of the mean of this "remainder" population. Moreover, we can express the means in the remainder sample and the remainder population in terms of the overall sample and population means and those of the subsample, distinguished by an argument (p) . Thus (40.46) gives an estimator

$$u_p = \frac{nm_y - pm_y(p)}{n-p} - a(p) \left\{ \frac{nm_x - pm_x(p)}{n-p} - \frac{N\mu_x - pm_x(p)}{N-p} \right\},$$

which will be unbiased for $\{N\mu_y - pm_y(p)\}/(N-p)$. Thus the unbiased estimator of μ_y itself is $\{(N-p)u_p + pm_y(p)\}/N$, which we write

$$t_p = \frac{N-p}{N} \frac{n}{n-p} \{m_y - a(p)(m_x - \mu_x)\} - \frac{N-n}{N} \frac{p}{n-p} \{m_y(p) - a(p)(m_x(p) - \mu_x)\}. \quad (40.47)$$

The choice of an integer p is arbitrary in $1 \leq p \leq n-1$. For given p , the function $a(p)$ of the subsample is also arbitrary. We therefore have a large class of unbiased estimators of μ_y which make use of our knowledge of μ_x .

Exactly the same argument holds in the multivariate situation where x is a vector.

40.14 An undesirable feature of the general class of estimators (40.47) is that they depend on the order in which the sample is drawn. We can overcome this by considering t_p for every one of the $n!$ possible orderings of the sample and averaging to obtain \bar{t}_p —this average sometimes takes a simple form requiring little computation from the sample. (This averaging process is exactly the same as we carried out in 39.11 for similar reasons, although there the results were not computationally simple because sampling was with unequal probabilities.) Exercise 39.30, which now simplifies since we are sampling with equal probabilities, shows that the averaged estimator \bar{t}_p has variance which is never greater than that of any single t_p .

40.15 If in (40.47) we choose $a(p) = m_{y/x}(p) = \frac{1}{p} \sum_{i=1}^p \frac{y_i}{x_i}$ and $p = 1$, it reduces to

$$t_1 = \mu_x \frac{y_1}{x_1} + \frac{N-1}{N} \cdot \frac{n}{n-1} \left(m_y - \frac{y_1}{x_1} m_x \right)$$

where y_1, x_1 are the values of the first observation drawn. Averaging over all $n!$ orderings of the sample, we obtain

$$\bar{t}_1 = \mu_x m_{y/x} + \frac{N-1}{N} \frac{n}{n-1} (m_y - m_{y/x} m_x), \quad (40.48)$$

which is identical with $\hat{\mu}'_y$ defined at (40.7). What is more, if we choose any other value of p and the same $a(p)$ as above, the average value \bar{t}_p will be the same as (40.48),

though of course t_p itself will differ. Thus if $a(p) = m_{y/x}(p)$, we get an unbiased estimator based on $m_{y/x}$.

This result encourages us to look for an exactly unbiased version of the type (40.1) by putting $a(p) = m_y(p)/m_x(p)$. (40.47) becomes

$$t_p = \mu_x \frac{m_y(p)}{m_x(p)} + \frac{(N-p)}{N} \frac{n}{(n-p)} \left(m_y - \frac{m_y(p)}{m_x(p)} m_x \right),$$

which when averaged gives a similar form with $m_y(p)/m_x(p)$ replaced by its average. If $p = 1$, of course, it reduces to (40.48), since $a(p)$ is then exactly what we had previously. The next simplest choice is $p = n-1$, when the average value of $m_y(p)/m_x(p)$ over all permutations is seen to be $\frac{1}{n} \sum_{i=1}^n \left(\frac{nm_y - y_i}{nm_x - x_i} \right) = R$. Thus

$$t_{n-1} = \mu_x R + \frac{(N-p)}{N} \frac{n}{(n-p)} (m_y - R m_x) \quad (40.49)$$

is an unbiased estimator of μ_y .

40.16 Turning now to regression estimators, it is natural to investigate the choice $a(p) = b_{yx}(p)$. (40.37) reduces to

$$t_p = m_y - b_{yx}(p)(m_x - \mu_x) - \frac{(N-n)}{N} \frac{p}{(n-p)} \{ (m_y(p) - m_y) - b_{yx}(p)(m_x(p) - m_x) \}. \quad (40.50)$$

Averaging simply replaces $b_{yx}(p)$ by its average. $p = 1$ is now impossible, since b_{yx} is then nugatory. As before, $p = n-1$ is the next simplest, involving the calculation of the regression coefficient n times, omitting each of the observations in turn, and averaging to obtain $\bar{b}_{yx}(n-1)$. (40.50) reduces to

$$t_{n-1} = m_y - \bar{b}_{yx}(n-1)(m_x - \mu_x) - \frac{N-n}{Nn} \left\{ \sum_{i=1}^n b_{yx}(n-1) x_i - n \bar{b}_{yx}(n-1) m_x \right\}, \quad (40.51)$$

where x_i in the summation is the value omitted in calculating the $b_{yx}(n-1)$ which it multiplies. (40.51) is equivalent to the usual regression estimator (40.41) if all $b_{yx}(n-1)$ are the same, but not in general otherwise. However, when n is large, the $b_{yx}(n-1)$ can vary very little, and the estimators differ correspondingly little.

Estimation of variance

40.17 The sampling variance of the unbiased estimators (40.47) cannot be generally investigated, since everything depends upon the choice of $a(p)$. However, if we modify the estimation scheme slightly, we can at once obtain estimators whose variance can be estimated.

Suppose that the n observations are split into k subsamples as they are drawn, the r th subsample containing n_r observations, $\sum_{r=1}^k n_r = n$. We write the partial sum

$\sum_{r=1}^q n_r = n_{+q}$, so that $n_{+1} = n_1$ and $n_{+k} = n$. We re-label the estimator (40.47) as $t(p, n)$ to signify that a subsample of p is used in a sample of size n . Consider the sequence of $(k-1)$ estimators $t(n_{+1}, n_{+2})$, $t(n_{+2}, n_{+3}) \dots t(n_{+(k-1)}, n_{+k})$, in which each

estimator uses the complete sample of the previous estimator as subsample. These estimators are all uncorrelated, for the operation of taking expectations may be split into a sequence of k conditional expectations (cf. 39.35) corresponding to the divisions into the k subsamples, and for $r < s$

$$\begin{aligned} & E\{t(n_{+r}, n_{+(r+1)}), t(n_{+s}, n_{+(s+1)})\} \\ &= E \dots E \{t(n_{+r}, n_{+(r+1)})\}_{r+2} \dots E \{t(n_{+s}, n_{+(s+1)})\}_k \\ &= E \dots E \{t(n_{+r}, n_{+(r+1)})\mu_y\}_{r+1} = \mu_y^2. \end{aligned}$$

Thus we may use the result of Exercise 39.12 to estimate the variance of the mean of the sequence of k unbiased uncorrelated estimators $t(n_{+r}, n_{+(r+1)})$. The k estimators themselves need not be of the same form.

40.18 In 40.17 we did not require that the k subsamples be of equal size. We now suppose that they are, so that $n_r = n/k$. If we use each of the subsamples in turn to evaluate a particular $a(p) = a\left(\frac{n}{k}\right)$ in (40.47), and calculate $t\left(\frac{n}{k}, n\right)$ each time, its k values will no longer be uncorrelated as in 40.17. Their mean is

$$\begin{aligned} T\left(\frac{n}{k}, n\right) &= \frac{1}{k} \sum^k t\left(\frac{n}{k}, n\right) \\ &= m_y - \bar{a}\left(\frac{n}{k}\right)(m_x - \mu_x) + \frac{N-n}{Nk(k-1)} \sum^k \left\{a\left(\frac{n}{k}\right) - \bar{a}\left(\frac{n}{k}\right)\right\} \left\{m_x\left(\frac{n}{k}\right) - m_x\right\}, \end{aligned} \quad (40.52)$$

where a bar denotes averaging over the k values obtained. The first two terms on the right of (40.52) are precisely of the form (40.46), but are not unbiased because $\bar{a}\left(\frac{n}{k}\right)$ is calculated from the same sample as m_x ; their expectation is $\mu_y - C\left\{\bar{a}\left(\frac{n}{k}\right), m_x\right\}$. The last term in (40.52) is evidently an unbiased estimator of this covariance if the population is regarded as consisting of $\frac{N}{(n/k)}$ groups of size n/k , of which k are selected at random for the sample.

Modification of sampling scheme to eliminate bias

40.19 The whole of our discussion of ratio and regression estimators in this chapter has been concerned with modifying the form of the estimator to eliminate or reduce bias in equal-probabilities sampling without replacement. Another way of achieving these objectives is to change the sampling scheme so that the original estimators are rendered unbiased.

(39.20) shows that an unbiased estimator of μ_y is given by $\frac{1}{N} \sum_{i=1}^n \frac{y_i}{\pi_i}$ for any set of selection probabilities π_i . Suppose, then, that we choose

$$\pi_i = nx_i / \sum_{i=1}^N x_i, \quad (40.53)$$

where we must now assume the auxiliary variable x to be positive so that (40.53) is a set of probabilities. It then follows at once that $\mu_x m_{y/x}$ will be an unbiased estimator of μ_y . In other words, for this sampling scheme, $\hat{\mu}_y$ defined at (40.2) is exactly unbiased. On the other hand, if we regard the set of $\binom{N}{n}$ possible samples as the population from which one member is to be drawn, and work in terms of the means m_y, m_x of these samples, the same argument shows that if we make the single selection with

$$\pi_i = (m_x)_i / \sum_{i=1}^{\binom{N}{n}} (m_x)_i = (m_x)_i / \left\{ \binom{N}{n} \mu_x \right\}, \quad (40.54)$$

the estimator (39.20) becomes

$$\frac{1}{\binom{N}{n}} \frac{m_y}{m_x / \left\{ \binom{N}{n} \mu_x \right\}} = \mu_x m_y / m_x,$$

so that $\tilde{\mu}_y$ defined at (40.1) becomes exactly unbiased, as first observed by Lahiri (1951).

The variances of these estimators, and unbiased estimates of them, are obtainable as usual from (39.23-4). In the case of $\tilde{\mu}_y$, of course, at least two samples (or random subsamples of one sample) are necessary for variance estimation to be possible.

Nanjamma *et al.* (1959) discuss the general problem of modifying the sampling scheme to render ratio estimators unbiased, with applications to several types of survey design. See also Pathak (1964a).

Stratified and multi-stage sampling

40.20 Any ratio or regression estimator may be applied separately within each of a number of strata, provided that the population mean of x is known within each stratum. Alternatively, a single ratio or regression estimator may be applied using the combined results from all strata. We should expect the former procedure to be the more efficient in general. The details are given by Cochran (1963) for biased ratio and regression estimators.

Unbiased stratified ratio estimators are discussed by Nieto de Pascual (1961) and W. H. Williams (1961) in the univariate case and by Olkin (1958) for multivariate situations. Robson and Vithayasai (1961) consider a stratification-like situation where y and x can be expressed as the sum of k corresponding components. Kish and Hess (1959) derive asymptotic formulae for the variance of the biased combined ratio estimator in stratified multi-stage sampling.

40.21 Durbin (1953) pointed out that since, from (40.30),

$$\frac{m_y}{m_x} = \frac{\mu_y}{\mu_x} + \frac{m_y - \left(\frac{\mu_y}{\mu_x} \right) m_x}{\mu_x} + o(n^{-1/2}), \quad (40.55)$$

the ratio of sample means is asymptotically linear in $y - (\mu_y/\mu_x)x = z$, so that we have

$$\frac{m_y}{m_x} - \frac{\mu_y}{\mu_x} \doteq \frac{1}{n} \sum_{i=1}^n z_i / \mu_x. \quad (40.56)$$

It follows that for the estimation of $V\left(\frac{m_y}{m_x}\right)$ in multi-stage sampling, the discussion of 39.45–50 applies asymptotically and the Yates–Durbin rule of 39.46 and 39.49 may be used in large samples. This will also be true separately within each stratum of a multi-stage design. The same applies for regression estimators.

Two-phase sampling

40.22 Our discussions of stratified sampling (39.15–27), and of the use of an auxiliary variable to improve estimation efficiency in this chapter, both presupposed some knowledge of the population in order to make unbiased estimation possible. In the latter case, it was μ_x which had to be known, and in the former the relative sizes of the strata, N_l/N , which are required to define the estimator (39.38). If this essential information is not available, a procedure which sometimes suggests itself on practical grounds is to carry out a preliminary equal-probabilities random sample to obtain it, and follow this by the main sample devoted to the original purpose of estimating the population mean. Clearly, such a procedure will only be economically acceptable if the cost of the preliminary sample is small relative to the gain in efficiency achieved in the main sample as a result—we make this point more precise later.

A sampling scheme of this kind is called *two-phase sampling*. (We do not use the older name *double sampling*, which has already been used in Chapter 34 for a sequential method whose aim is to achieve a confidence interval of prescribed length and coefficient. This is certainly not our purpose here, where we aim primarily to improve efficiency of estimation at the second phase by collecting auxiliary information at the first phase.) Two-phase sampling is distinguished from two-stage sampling by the fact that it uses the same sampling units at each phase of sampling.

40.23 Following Neyman (1938), who first solved the problem, we consider the stratification problem first. We wish to stratify into a fixed number k of defined strata, but are ignorant of the population proportions $N_l/N = W_l$ in these strata and therefore cannot use the estimator (39.38). Accordingly we take a preliminary equal-probabilities sample of size n_1 , which is found to be distributed over the k desired strata with frequencies $n_{11}, n_{12}, \dots, n_{1k}$, where $\sum_{l=1}^k n_{1l} = n_1$. The proportions $w_{1l} = n_{1l}/n_1$ are, of course, unbiased estimators of the population proportions W_l , and it is therefore natural to use as our estimator of μ

$$\hat{\mu}_{12} = \sum_{l=1}^k w_{1l} m_{2l}, \quad (40.57)$$

where the m_{2l} are the sample stratum means in the second (main) sample, of size n_2 , with n_{2l} observations in the l th stratum. The question now arises whether n_{2l} is to be a subsample of n_{1l} or whether it is to be entirely independently selected. In practice, the former is much more likely, since if N_l is unknown no complete listing of the strata can be available and second-phase sampling will be based on the random sample in each stratum obtained at the first phase. Furthermore, although the first phase is

logically prior to the second, the two phases can sometimes be carried out simultaneously if the second uses a subsample of the first.

We assume that the n_{1l} are certain to be so large, in relation to the intended values of the n_{2l} , that the observed values of the n_{1l} are no restriction upon the fixing of the n_{2l} in advance.

40.24 We now (cf. 39.35) use the symbol E_2 to denote taking expectations at the second phase, conditional upon the first-phase results being fixed, and E_1 to denote expectations at the first phase. Using (39.72), the expectation of (40.57) is

$$\begin{aligned} E(\hat{\mu}_{12}) &= E_1 \{ E_2(\hat{\mu}_{12}) \} = \sum_l E_1 \{ w_{1l} E_2(m_{2l}) \} \\ &= \sum_l E_1 \{ w_{1l} \} \mu_l = \sum_l W_l \mu_l = \mu, \end{aligned} \quad (40.58)$$

so that $\hat{\mu}_{12}$ is unbiased. Its variance, from (39.73), is

$$V(\hat{\mu}_{12}) = E_1 \{ V_2(\hat{\mu}_{12}) \} + V_1 \{ E_2(\hat{\mu}_{12}) \}. \quad (40.59)$$

The first term on the right of (40.59) is evaluated by observing that just as at (39.39)

$$V_2(\hat{\mu}_{12}) = V_2(\sum_l w_{1l} m_{2l}) = \sum_l w_{1l}^2 \frac{\sigma_l^2}{n_{2l}} \left(1 - \frac{n_{2l}}{N_l} \right),$$

and hence

$$E_1 \{ V_2(\hat{\mu}_{12}) \} = \sum_l E_1(w_{1l}^2) \frac{\sigma_l^2}{n_{2l}} \left(1 - \frac{n_{2l}}{N_l} \right). \quad (40.60)$$

If we now assume each N_l to be very large, the last factor on the right of (40.60) is negligible. What is more, the first-phase sampling is now effectively multinomial estimation of proportions, so that (5.80) applies and (40.60) becomes

$$E_1 \{ V_2(\hat{\mu}_{12}) \} = \sum_l \left\{ \frac{W_l(1-W_l)}{n_1} + W_l^2 \right\} \frac{\sigma_l^2}{n_{2l}}. \quad (40.61)$$

The second term on the right of (40.59) is, using (10.16) and (5.80),

$$\begin{aligned} V_1 \{ E_2(\hat{\mu}_{12}) \} &= V_1 \{ \sum_l w_{1l} \mu_l \} = \sum_l \mu_l^2 V_1(w_{1l}) + \sum_{\substack{l, p \\ l \neq p}} \mu_l \mu_p C_1(w_{1l}, w_{1p}) \\ &= \sum_l \mu_l^2 \frac{W_l(1-W_l)}{n_1} - \sum_{\substack{l, p \\ l \neq p}} \mu_l \mu_p \frac{W_l W_p}{n_1} \\ &= \frac{1}{n_1} \{ \sum_l W_l \mu_l^2 - (\sum_l W_l \mu_l)^2 \}. \end{aligned} \quad (40.62)$$

Putting (40.61-2) into (40.59), we find, since $\sum_l W_l \mu_l = \mu$,

$$V(\hat{\mu}_{12}) = \sum_l \left\{ \frac{W_l(1-W_l)}{n_1} + W_l^2 \right\} \frac{\sigma_l^2}{n_{2l}} + \frac{1}{n_1} \sum_l W_l (\mu_l - \mu)^2. \quad (40.63)$$

It will be recalled from (39.46) and 39.19 that the last term on the right of (40.63) expresses the gain in precision of a USF stratified sample over an unstratified sample when all stratum sizes are large and n_1 is sample size.

40.25 If $n_1 \rightarrow \infty$, so that the W_l are effectively known, (40.63) reduces to the usual stratified sampling variance formula (39.39). Even for moderate n_1 , the term $W_l(1-W_l)/n_1$, which cannot exceed $1/(4n_1)$, will usually be negligible compared with W_l^2 . From 39.18, we recall that $\sum_l W_l(\mu_l - \mu)^2 \doteq \sigma^2 - \sum_l W_l \sigma_l^2$ for N_l large, so that

(40.63) may be approximated by

$$V(\hat{\mu}_{12}) \doteq \frac{\sigma^2 - \sum_l W_l \sigma_l^2}{n_1} + \sum_l W_l^2 \frac{\sigma_l^2}{n_{2l}}. \quad (40.64)$$

An almost unbiased estimator of (40.64) is obtained by substituting w_{1l} for W_l , s^2 for σ^2 , and s_l^2 for σ_l^2 .

40.26 Suppose now that the cost function for the two-phase sample is

$$C = c_0 + n_1 c_1 + \sum_{l=1}^k n_{2l} c_{2l}. \quad (40.65)$$

(40.64-5) are of the form (39.50-1). It follows from (39.53) that the sample sizes which minimize $V(\hat{\mu}_{12})$ for fixed C (or vice versa) are

$$\left. \begin{aligned} n_1^2 &\propto \frac{\sigma^2 - \sum_l W_l \sigma_l^2}{c_1}, \\ n_{2l}^2 &\propto \frac{W_l^2 \sigma_l^2}{c_{2l}}, \end{aligned} \right\} \quad (40.66)$$

the constant of proportionality being obtained by (40.65) or (40.64), whichever is fixed.

(40.66) shows that at the second phase, observations should be distributed between the strata just as in ordinary stratified sampling allocation at (39.55) (though it must be remembered that only the neglect of a term of order n_1^{-1} has produced this simple result). The first-phase sample size is directly proportional to the numerator of the first term on the right of (40.64) (which is the excess variance resulting from the need to estimate the W_l at the first phase) and inversely proportional to the cost of sampling, both considerations in accord with intuition.

40.27 Although the intention of our two-phase sampling is to improve estimation efficiency by use of stratification, we recall from 39.18 that even when the W_l are known precisely, the best stratification may cause a loss of efficiency, though we saw in 39.19 that this could not happen if all the N_l were large enough. However, the additional component of sampling variance due to the estimation of the W_l at the first-phase sampling now opens the possibility of a loss of efficiency even for large N_l .

Consider an equal-probabilities unstratified random sample of size n , with n_l observations falling into the l th stratum. It is reasonable to assume that the overhead cost of the sample will be the same c_0 as in (40.65), and that the cost of an observation in the l th stratum will also remain unchanged at c_{2l} . Thus the cost function is

$$C_R = c_0 + \sum_l n_l c_{2l}.$$

However, n_l is now a random variable with expectation nW_l , so that we must work with the expected cost

$$E(C_R) = c_0 + n \sum_l W_l c_{2l}. \quad (40.67)$$

If $E(C_R)$ is to equal C at (40.65), we must have

$$n = \frac{n_1 c_1 + \sum_l n_{2l} c_{2l}}{\sum_l W_l c_{2l}}$$

and thus the variance of the unstratified estimator will be (for large N)

$$V(m_R) = \frac{\sigma^2}{n} = \frac{\sigma^2 \sum_l W_l c_{2l}}{n_1 c_1 + \sum_l n_{2l} c_{2l}}. \quad (40.68)$$

The ratio of two-phase stratified to unstratified sampling variance is, from (40.64) and (40.68),

$$\frac{V(\hat{\mu}_{12})}{V(m_R)} = \frac{\left(\frac{\sigma^2 - \sum_l W_l \sigma_l^2}{n_1} + \sum_l W_l^2 \frac{\sigma_l^2}{n_{2l}} \right) (n_1 c_1 + \sum_l n_{2l} c_{2l})}{\sigma^2 \sum_l W_l c_{2l}}. \quad (40.69)$$

The numerator of (40.69) is the product $V(\hat{\mu}_{12})(C - c_0)$, which is minimized when n_1 and n_{2l} are chosen to satisfy (40.66). By (39.52), this minimum value is given by

$$\frac{\min V(\hat{\mu}_{12})}{V(m_R)} = \frac{[\{(\sigma^2 - \sum_l W_l \sigma_l^2) c_1\}^{\frac{1}{2}} + \sum_l W_l \sigma_l c_{2l}^{\frac{1}{2}}]^2}{\sigma^2 \sum_l W_l c_{2l}}. \quad (40.70)$$

This seems to be the most useful form for the ratio of variances. If we again consider the numerator (40.70), we see by the Cauchy inequality that it is no greater than

$$\{(\sigma^2 - \sum_l W_l \sigma_l^2) + \sum_l W_l \sigma_l^2\} \{c_1 + \sum_l W_l c_{2l}\}$$

so that (40.70) gives

$$\frac{\min V(\hat{\mu}_{12})}{V(m_R)} \leq 1 + \frac{c_1}{\sum_l W_l c_{2l}}. \quad (40.71)$$

Thus if $c_1 = 0$, two-phase stratified sampling with MV allocation of sample sizes is never worse than unstratified sampling with the same expected cost. But if $c_1 = 0$, we can estimate the W_l accurately at zero cost, so this is effectively ordinary stratified sampling. We have thus verified the conclusion of 39.19 with the additional consideration of variable costs in the different strata.

If $c_1 > 0$ in (40.70), it is possible for the unstratified sample to be more efficient, but (40.70) is evidently an increasing function of c_1 , and if c_1 is small compared to the weighted average $\sum_l W_l c_{2l}$, the right-hand side of (40.71) can exceed unity by very little, so that there is, at worst, little efficiency to be lost by properly allocated two-phase stratified sampling. As a simple unfavourable numerical example, put $c_1 = 1$, $c_{2l} = 6$, all l , $\sigma^2 = 10$, $\sigma_l^2 = 6$, all l . The value of (40.70) is then $(2 + 6)^2 / (10 \times 6) = 1.07$. If instead $c_{2l} = 9$, (40.70) becomes $(1 + 7.35)^2 / (10 \times 9) = 0.77$.

40.28 When the first phase of sampling is being used to estimate the mean of an auxiliary variable, μ_x , for a ratio or regression estimator at the second phase, (39.73) is used to evaluate the sampling variance, as in 40.24. We shall consider only the simplest case, using the biased ratio estimator (40.1). In two-phase sampling this is

$$\tilde{\mu}_{12} = m_x^{(1)} \cdot m_y^{(2)} / m_x^{(2)},$$

where we now use superscripts to denote phases. If the two phases are independent equal-probabilities samples, we have, using 40.2,

$$\begin{aligned} E(\tilde{\mu}_{12}) &= E_1 \{m_x^{(1)} E_2(m_y^{(2)}/m_x^{(2)})\} = E_1 \{m_x^{(1)} [\mu_y/\mu_x + O(n_2^{-1})]\} \\ &= \mu_y + O(n^{-1}). \end{aligned}$$

(39.73) gives

$$\begin{aligned} V(\tilde{\mu}_{12}) &= E_1 \{V_2(\tilde{\mu}_{12})\} + V_1 \{E_2(\tilde{\mu}_{12})\} \\ &= E_1 \{(m_x^{(1)})^2 V_2(m_y^{(2)}/m_x^{(2)})\} + V_1 \{m_x^{(1)} \mu_y/\mu_x\}, \end{aligned} \quad (40.72)$$

where we neglect terms of relative order n_2^{-1} . Thus

$$V(\tilde{\mu}_{12}) \doteq V_2(m_y^{(2)}/m_x^{(2)}) \{V_1(m_x^{(1)}) + \mu_x^2\} + \left(\frac{\mu_y}{\mu_x}\right)^2 V(m_x^{(1)}),$$

and using (40.24), this becomes

$$V(\tilde{\mu}_{12}) \doteq \frac{1}{n_2} \left\{ V(y) + \left(\frac{\mu_y}{\mu_x}\right)^2 V(x) - 2\frac{\mu_y}{\mu_x} C(y, x) \right\} \left(1 + \frac{V(x)}{n_1 \mu_x^2}\right) + \left(\frac{\mu_y}{\mu_x}\right)^2 \frac{V(x)}{n_1}. \quad (40.73)$$

The term in $1/n_2$ on the right of (40.73) is simply (40.25) applied to the second-phase sample. As in (40.63), the first-phase sampling introduces a term in $1/n_1$ inflating this contribution, as well as a new contribution of order $1/n_1$. Since we have already neglected terms of relative order $1/n_2$, we also (since we assume $n_1 > n_2$) neglect the term in $1/n_1 n_2$, obtaining the approximation

$$V(\tilde{\mu}_{12}) \doteq \frac{1}{n_2} \left\{ V(y) + \left(\frac{\mu_y}{\mu_x}\right)^2 V(x) - 2\frac{\mu_y}{\mu_x} C(y, x) \right\} + \left(\frac{\mu_y}{\mu_x}\right)^2 \frac{V(x)}{n_1}. \quad (40.74)$$

If, instead of being independent, the second-phase sample is a subsample of the first, (40.72) is modified, μ_y/μ_x there being replaced by $m_y^{(1)}/m_x^{(1)}$, so that the second term on its right-hand side becomes simply $V(m_y^{(1)}) = V(y)/n_1$. If n_1 is very much larger than n_2 , the first term on the right of (40.72) has the same value as previously to our order of approximation, but if n_2/n_1 is appreciable, the approximation is improved by a correction $\left(1 - \frac{n_2}{n_1}\right)$ applied to the first term in (40.74), which then becomes

$$V(\tilde{\mu}_{12}) \doteq \left(\frac{1}{n_2} - \frac{1}{n_1}\right) \left\{ V(y) + \left(\frac{\mu_y}{\mu_x}\right)^2 V(x) + 2\frac{\mu_y}{\mu_x} C(y, x) \right\} + \frac{V(y)}{n_1}. \quad (40.75)$$

40.29 Cochran (1963) and Yates (1960) give details of application of two-phase methods to regression estimators, although restrictive assumptions are necessary to obtain useful variance formulae.

Two-phase sampling generalizes naturally to *multi-phase sampling*, but little theoretical or practical work has been carried out on this more general procedure.

S. P. Ghosh (1963a) considers a form of two-phase sampling where the object of the first phase is to form clusters for the second-phase sampling.

Raj (1964) discusses the case where the first phase is used to determine probabilities of selection to be used at the second phase.

Domains of study

40.30 In our discussions of ratio estimation, we have allowed the auxiliary variable, x , to be quite general. In practice, one of the most important situations is that in which x is a 0-1 variable which counts whether the corresponding value of y is not, or is, a member of a certain sub-group of the population. Among the situations falling into this class are the following:

- (a) A population is sampled, but we are from the outset only interested in part of it; e.g. the human population aged 21 and over is sampled, but we are interested only in the ages 21-65. Here the sample size n for the population of interest is clearly a random variable, and the sample mean for any variable measured in this population is of the form $\sum_{i=1}^n y_i / \sum_{i=1}^n x_i$, where x_i is 0 or 1 as above. If the population mean of x (i.e. the proportion of the population aged 21-65) is known, all our foregoing theory can be applied.
- (b) We are interested in the entire population from which we sample, but only part of the selected sample yields observations, owing to non-response (in human populations, especially), loss of records, or incomplete fieldwork. Again, n is a random variable, and the remarks under (a) apply. (*)
- (c) We obtain observations from the whole sample taken from the population of interest, but we wish to evaluate the results for sub-groups of this population; e.g. we have a sample from a human population, and wish to calculate certain statistics for men only and for women only. If we had stratified the sample in advance into men and women, no new point would have arisen, since sample sizes for men and for women would be fixed. However, such stratification is not usually possible, so that these sample sizes are random variables (though their sum is not, in this simple case). More generally, the sample size in any unpredesignated sub-group must be a random variable.

40.31 The sub-group of interest is called a *domain* (of study). Of course, a stratum may itself be a domain, but no new theory is then required. We shall use domain to mean a sub-group whose sample frequency is a random variable, whatever the reason.

Domains frequently cut across the strata and the various stage-units of a sample, and it is here that new points arise. Yates (1960) gives (as also in earlier editions of his book) a number of formulae for domains cutting across strata, for which Cochran (1963) gives some of the derivations. Durbin (1958) treats these and some multi-stage situations. Hartley (1959) also derives some of Yates' results for covariances of domain means, and gives some further results. Our treatment follows Durbin (1958).

Domains across strata

40.32 Suppose, as in Chapter 39, that we have k strata with population frequencies N_l , $l = 1, 2, \dots, k$, and $\sum_l N_l = N$, while in the sample the stratum frequencies are

(*) There is a complication here in the case of non-response, since non-response may be correlated with the value of y , so that the responding group cannot provide an unbiased estimator of y for the population as a whole.

n_l with $\sum_l n_l = n$. Now consider a particular domain, say d . We denote the population frequency in d in the l th stratum by $N_l^{(d)}$ with $\sum_l N_l^{(d)} = N^{(d)}$ for the domain frequency in the entire population, while $n_l^{(d)}$ and $\sum_l n_l^{(d)} = n^{(d)}$ are similarly defined in the sample. Note that whereas the population stratum frequencies N_l are known, the population stratum domain frequencies $N_l^{(d)}$ will generally not be known. Of course, unstratified random sampling is the case $k = 1$.

We define the variable $y^{(d)}$ to be equal to the observed variable y for domain members, and to be zero for others. Thus

$$y_{lj}^{(d)} = h_{lj} y_{lj}, \quad (40.76)$$

where

$$h_{lj} = \begin{cases} +1 & \text{within the domain,} \\ 0 & \text{outside the domain.} \end{cases} \quad (40.77)$$

We then have

$$\left. \begin{aligned} N_l^{(d)} &= \sum_{j=1}^{N_l} h_{lj}, & N^{(d)} &= \sum_{l=1}^k N_l^{(d)} = \sum_{l=1}^k \sum_{j=1}^{N_l} h_{lj}, \\ n_l^{(d)} &= \sum_{j=1}^{n_l} h_{lj}, & n^{(d)} &= \sum_{l=1}^k n_l^{(d)} = \sum_{l=1}^k \sum_{j=1}^{n_l} h_{lj}. \end{aligned} \right\} \quad (40.78)$$

We further define the domain means within strata,

$$\mu_l^{(d)} = \sum_{j=1}^{N_l} y_{lj}^{(d)} / N_l^{(d)}, \quad (40.79)$$

and the overall domain mean

$$\mu^{(d)} = \sum_{l=1}^k \sum_{j=1}^{N_l} y_{lj}^{(d)} / N^{(d)} = \sum_l N_l^{(d)} \mu_l^{(d)} / N^{(d)}. \quad (40.80)$$

40.33 We now seek to estimate $\mu^{(d)}$ at (40.80). Consider first the case where the sampling is with equal probabilities using a USF, say $f = n/N$ as in Chapter 39. The ratio estimator

$$m^{(d)} = \sum_l \sum_{j=1}^{n_l} y_{lj}^{(d)} / n^{(d)} = \sum_l \sum_j y_{lj}^{(d)} / \sum_l \sum_j h_{lj} \quad (40.81)$$

is the sample analogue of $\mu^{(d)}$. It is in essence (40.1) with numerator and denominator separately summed across strata—i.e. it is an example of the “combined” ratio estimator referred to in 40.20. Using the analogues of (40.3) and (40.5), we see that to the first order of approximation

$$\begin{aligned} E(m^{(d)}) &\sim E\left(\sum_l \sum_{j=1}^{n_l} y_{lj}^{(d)}\right) / E\left(\sum_l \sum_{j=1}^{n_l} h_{lj}\right) \\ &= \left(\sum_l f \sum_{j=1}^{N_l} y_{lj}^{(d)}\right) / \left(\sum_l f \sum_{j=1}^{N_l} h_{lj}\right) \\ &= \mu^{(d)}, \end{aligned} \quad (40.82)$$

using (40.78) and (40.80).

40.34 To find the variance of $m^{(d)}$, we put

$$z_{lj} = h_{lj}(y_{lj} - \mu^{(d)}) = y_{lj}^{(d)} - h_{lj}\mu^{(d)}, \quad (40.83)$$

so that, using (40.78) and (40.81),

$$(40.84) \text{ is asymptotically } m^{(d)} - \mu^{(d)} = \sum_l \sum_{j=1}^{n_l} z_{lj} / n^{(d)}. \quad (40.84)$$

$$m^{(d)} - \mu^{(d)} \sim \sum_l \sum_j z_{lj} / E(n^{(d)}) = \sum_l \sum_j z_{lj} / \left(n \frac{N^{(d)}}{N} \right). \quad (40.85)$$

Thus

$$V(m^{(d)}) \sim \left(\frac{N}{N^{(d)}} \right)^2 V \left(\frac{1}{n} \sum_l \sum_{j=1}^{n_l} z_{lj} \right). \quad (40.86)$$

The variance on the right of (40.86) is that of a mean in a USF stratified sample. By (39.45), therefore, we have

$$\begin{aligned} V(m^{(d)}) &\sim \left(\frac{N}{N^{(d)}} \right)^2 \frac{N-n}{nN^2} \sum_l N_l \sigma_l^2(z) \\ &= \frac{N-n}{n(N^{(d)})^2} \sum_l \left[\sum_{j=1}^{N_l} z_{lj}^2 - \frac{\left(\sum_{j=1}^{N_l} z_{lj} \right)^2}{N_l} \right], \end{aligned} \quad (40.87)$$

where we have ignored the fact that $\sigma_l^2(z)$ has $N_l - 1$ as divisor rather than N_l . From (40.83), (40.87) may be written

$$\begin{aligned} \frac{n(N^{(d)})^2}{N-n} V(m^{(d)}) &\sim \sum_l \left[\sum_{j=1}^{N_l} \{ h_{lj} (y_{lj} - \mu^{(d)}) \}^2 - \frac{\left\{ \sum_{j=1}^{N_l} h_{lj} (y_{lj} - \mu^{(d)}) \right\}^2}{N_l} \right] \\ &= \sum_l \left[\sum_{j=1}^{N_l} h_{lj} (y_{lj} - \mu^{(d)})^2 - \frac{\{ N_l^{(d)} (\mu_l^{(d)} - \mu^{(d)}) \}^2}{N_l} \right] \end{aligned} \quad (40.88)$$

using (40.78-9) and the fact that $h_{lj}^2 = h_{lj}$. The effect of the term h_{lj} in the first summation over j on the right of (40.88) is to convert the y_{lj} in the succeeding parenthesis to $y_{lj}^{(d)}$ by (40.76), and to leave $\mu^{(d)}$ there unchanged, since by (40.80) this is a linear function of the $y_{lj}^{(d)}$ and $h_{lj} y_{lj}^{(d)} = y_{lj}^{(d)}$. This first summation may therefore be written

$$\begin{aligned} \sum_{j=1}^{N_l} h_{lj} (y_{lj} - \mu^{(d)})^2 &= \sum_j (y_{lj}^{(d)} - \mu^{(d)})^2 \\ &= \sum_j (y_{lj}^{(d)} - \mu_l^{(d)})^2 + N_l^{(d)} (\mu_l^{(d)} - \mu^{(d)})^2, \end{aligned} \quad (40.89)$$

by the usual sum of squares identity. Putting (40.89) into (40.88), we obtain

$$V(m^{(d)}) \sim \frac{N-n}{n(N^{(d)})^2} \sum_l \left[\sum_{j=1}^{N_l} (y_{lj}^{(d)} - \mu_l^{(d)})^2 + N_l^{(d)} \left(1 - \frac{N_l^{(d)}}{N_l} \right) (\mu_l^{(d)} - \mu^{(d)})^2 \right]. \quad (40.90)$$

40.35 The first term on the right of (40.90) is, if we write $n/N = n^{(d)}/N^{(d)}$ to our order of approximation,

$$\frac{N^{(d)} - n^{(d)}}{n^{(d)} (N^{(d)})^2} \sum_l \sum_j (y_{lj}^{(d)} - \mu_l^{(d)})^2. \quad (40.91)$$

As is evident from our derivation of (40.87) from (39.45), (40.91) is exactly the "variance" we should have arrived at if the stratum domain frequencies $n_l^{(d)}$ had been (wrongly) regarded as fixed. The second term in (40.90) therefore indicates the increase in variance attributable to variation in the $n_l^{(d)}$. This will be large if the

stratum domain means $\mu_i^{(d)}$ differ substantially, particularly if the fractions of the strata frequencies within the domain, $N_i^{(d)}/N_i$, are small.

If we drop the factor $\left(1 - \frac{N_i^{(d)}}{N_i}\right)$ in the second term on the right of (40.90), and use the approximation in (40.91), the whole of (40.90) is to our degree of approximation identical with the first formula of 39.18, which will be seen to be the *unstratified* sampling variance. Thus if all the fractions $N_i^{(d)}/N_i$ are small, we see that the variation in the $n_i^{(d)}$ effectively removes the whole benefit of stratification from the sampling variance of the estimator. Only if the domain bulks large within at least some of the strata is much of the benefit retained.

An estimator of the sampling variance (40.90) can be derived in exactly the same way as (40.90) itself was—the details are left to the reader as Exercise 40.15.

40.36 If the sampling fraction is not uniform, the estimator (40.81) must be changed to weight the stratum contributions properly. Instead, we put

$$r = \sum_i \frac{N_i}{n_i} \sum_{j=1}^{n_i} y_{ij} / \sum_i \frac{N_i}{n_i} n_i^{(d)}. \quad (40.92)$$

The reader is asked to show in Exercise 40.16 that this is asymptotically unbiased for $\mu^{(d)}$, with variance

$$V(r) \sim \frac{1}{(N^{(d)})^2} \sum_i \frac{N_i}{n_i} \left(1 - \frac{n_i}{N_i}\right) \sum_{j=1}^{n_i} \left(z_{ij} - \frac{\sum_{j=1}^{N_i} z_{ij}}{N_i} \right)^2, \quad (40.93)$$

the generalization of (40.87), which reduces on substitution for z_{ij} to

$$V(r) = \left(\sum_i \frac{N_i}{n_i} n_i^{(d)} \right)^{-2} \sum_i \frac{N_i}{n_i} \left(1 - \frac{n_i}{N_i}\right) \left[\sum_{j=1}^{N_i} (y_{ij}^{(d)} - \mu_i^{(d)})^2 + N_i^{(d)} \left(1 - \frac{N_i^{(d)}}{N_i}\right) (\mu_i^{(d)} - \mu^{(d)})^2 \right] \quad (40.94)$$

which generalizes (40.90). An estimator of (40.94) is

$$\hat{V}(r) = \left(\sum_i \frac{N_i}{n_i} n_i^{(d)} \right)^{-2} \sum_i \frac{N_i^2}{n_i(n_i-1)} \left(1 - \frac{n_i}{N_i}\right) \left[\sum_{j=1}^{n_i} (y_{ij}^{(d)} - m_i^{(d)})^2 + n_i^{(d)} \left(1 - \frac{n_i^{(d)}}{n_i}\right) (m_i^{(d)} - m^{(d)})^2 \right], \quad (40.95)$$

the derivation again being left to Exercise 40.16.

Domains in multi-stage sampling

40.37 We shall confine our attention to multi-stage sampling in which s first-stage units are selected with replacement from S units. Any number (including zero) of stages of selection is permitted thereafter, but we restrict ourselves further to self-weighting designs (cf. 39.41), for which the sample mean is the estimator of the corresponding population mean μ .

We wish to estimate the overall domain mean, written $\mu^{(d)}$ as before, where

$$\mu^{(d)} = \sum_{i=1}^S \sum_j \dots \sum_p y_{ij}^{(d)} \dots_p / N^{(d)} \quad (40.96)$$

and $N^{(d)}$ is the total domain frequency as before,

$$N^{(d)} = \sum_{i=1}^s \sum_j \dots \sum_p h_{ij} \dots_p = \sum_{i=1}^s N_i^{(d)}, \quad (40.97)$$

the $h_{ij} \dots_p$ being 0-1 variables as at (40.77).

40.38 The estimator is the sample analogue of $\mu_j^{(d)}$,

$$m^{(d)} = \sum_{i=1}^s \sum_j \dots \sum_p y_{ij}^{(d)} \dots_p / n^{(d)} \equiv \sum_{i=1}^s y_i^{(d)} / n^{(d)}, \quad (40.98)$$

where

$$n^{(d)} = \sum_{i=1}^s \sum_j \dots \sum_p h_{ij} \dots_p \equiv \sum_{i=1}^s n_i^{(d)}. \quad (40.99)$$

As at (40.82), to the first order of approximation,

$$\begin{aligned} E(m^{(d)}) &\sim E\left(\sum_{i=1}^s \sum_j \dots \sum_p y_{ij}^{(d)} \dots_p\right) / E(n^{(d)}) \\ &= f \sum_{i=1}^s \sum_j \dots \sum_p y_{ij}^{(d)} \dots_p / f \sum_{i=1}^s \sum_j \dots \sum_p h_{ij} \dots_p \\ &= \mu^{(d)}, \end{aligned} \quad (40.100)$$

the common factor f in numerator and denominator being the overall probability of selection for each value y in the population.

40.39 Just as at (40.83), we define

$$z_{ij} \dots_p = y_{ij}^{(d)} \dots_p - h_{ij} \dots_p \mu^{(d)}, \quad (40.101)$$

and find as at (40.84) that

$$\begin{aligned} m^{(d)} - \mu^{(d)} &= \frac{1}{n^{(d)}} \sum_{i=1}^s \sum_j \dots \sum_p z_{ij} \dots_p \\ &= \frac{s}{n^{(d)}} \cdot \frac{1}{s} \sum_{i=1}^s (\sum_j \dots \sum_p z_{ij} \dots_p). \end{aligned} \quad (40.102)$$

Thus, proceeding immediately to the estimation of the variance of $m^{(d)}$, we have

$$\hat{V}(m^{(d)}) \sim \left\{ \frac{s}{E(n^{(d)})} \right\}^2 \hat{V}(\bar{z}) \quad (40.103)$$

where \bar{z} is the mean of the s values $\sum_j \dots \sum_p z_{ij} \dots_p = z_i$, say. Now from (39.109),

$$\hat{V}(\bar{z}) = \hat{V}_1(\bar{z})$$

and in particular, if the probabilities of selection are the same at each first-stage drawing,

(39.110) gives in this case, with $t_i = \frac{z_i}{s}$,

$$\hat{V}(\bar{z}) = \hat{V}_1(\bar{z}) = \frac{1}{s(s-1)} \sum_{i=1}^s (z_i - \bar{z})^2.$$

Thus (40.103) becomes

$$\hat{V}(m^{(d)}) \sim \frac{s}{(s-1) \{E(n^{(d)})\}^2} \sum_{i=1}^s (z_i - \bar{z})^2. \quad (40.104)$$

Just as at (40.87-8), we now resolve the sum of squares in (40.104), using (40.101), into

$$\begin{aligned} \sum_{i=1}^s (z_i - \bar{z})^2 &= \sum_i z_i^2 - s\bar{z}^2 \\ &= \sum_i (y_i^{(d)} - n_i^{(d)} \mu^{(d)})^2 - \frac{(n^{(d)})^2}{s} (m^{(d)} - \mu^{(d)})^2, \end{aligned} \quad (40.105)$$

using the definitions of $y_i^{(d)}$ and $n_i^{(d)}$ in (40.98-9). (40.104-5) give, assuming that $n^{(d)} \sim E(n^{(d)})$,

$$\hat{V}(m^{(d)}) \sim \frac{s}{(s-1)(n^{(d)})^2} \sum_{i=1}^s (y_i^{(d)} - n_i^{(d)} \mu^{(d)})^2 - \frac{1}{s-1} (m^{(d)} - \mu^{(d)})^2, \quad (40.106)$$

and the expectation of the second term on the right of (40.106) is $-\frac{1}{s-1} V(m^{(d)})$ to our order of approximation. Thus, taking it to the left-hand side, we find

$$\hat{V}(m^{(d)}) \sim \frac{1}{(n^{(d)})^2} \sum_{i=1}^s (y_i^{(d)} - n_i^{(d)} \mu^{(d)})^2. \quad (40.107)$$

(40.107) is still not a statistic, as it depends on $\mu^{(d)}$. To our order of approximation this may be replaced by $m^{(d)}$, so that finally we obtain the estimator

$$\hat{V}(m^{(d)}) = \frac{1}{(n^{(d)})^2} \sum_{i=1}^s (y_i^{(d)} - n_i^{(d)} m^{(d)})^2. \quad (40.108)$$

If there is no sampling after the first stage, (40.108) agrees to this order of approximation with the result of Exercise 40.15.

40.40 If we had (wrongly) taken the $n_i^{(d)}$ to be fixed, we should have found for the estimated "variance" of (40.98), from (39.110) with $t_i = y_i^{(d)}/n^{(d)}$,

$$\frac{1}{(n^{(d)})^2} \frac{s}{s-1} \sum_{i=1}^s \left(y_i^{(d)} - \frac{n^{(d)}}{s} m^{(d)} \right)^2. \quad (40.109)$$

Comparison of (40.108) and (40.109) shows that the variation in the $n_i^{(d)}$ affects the variance by replacing the average domain frequency in a first-stage unit, $n^{(d)}/s$, by the individual first-stage unit domain frequencies, $n_i^{(d)}$. The increase in variance will be large only if the $n_i^{(d)}$ vary substantially and if they are negatively correlated with the $y_i^{(d)}$, which is unlikely in practice.

EXERCISES

40.1 Show from (40.7) and (40.9) that

$$\frac{\mu_x}{n-1} (nm_y/m_x - m_{y/x})$$

is an approximately unbiased estimator of μ_y .

(cf. M. N. Murthy and Nanjamma, 1959)

40.2 Show that the product estimator

$$\bar{\mu}_y = m_y m_x / \mu_x,$$

analogous to the ratio estimator (40.1), has bias $C(m_y, m_x)/\mu_x$ and hence that

$$\bar{\mu}'_y = \left\{ n(N-1)m_y m_x - (N-n) \sum_{i=1}^n y_i x_i / n \right\} / \{N(n-1)\mu_x\}$$

is unbiased. Show further that, as $N \rightarrow \infty$,

$$V(\bar{\mu}'_y) = \frac{\mu_y^2}{n} \left[\left(\frac{V(y)}{\mu_y^2} + \frac{V(x)}{\mu_x^2} + \frac{2C(y, x)}{\mu_y \mu_x} \right) + \frac{1}{n-1} \left(\frac{V(y)V(x) + \{C(y, x)\}^2}{\mu_y^2 \mu_x^2} \right) \right].$$

Hence show that if $C(y, x) < 0$, $\bar{\mu}'_y$ is more efficient than $\bar{\mu}_y'$, whose mean-square error is given at (40.25); while if $\beta_{yx} = \frac{C(y, x)}{V(x)} < -\frac{1}{2} \frac{\mu_y}{\mu_x}$, $\bar{\mu}'_y$ is more efficient than the simple mean m_y .

(Robson, 1957)

40.3 In inverse sampling (cf. Examples 9.13, 34.1) of a population of N individuals with unequal probabilities and with replacement, sampling continues until $(r+1)$ distinct individuals have been selected, when $(n+1)$ observations will have been made ($n \geq r$). The last observation is ignored, leaving r distinct values y_i , observed n_i times respectively, with $\sum_{i=1}^r n_i = n$.

Show that $t = \frac{1}{n} \sum_{i=1}^r n_i y_i$ is an unbiased estimator of the population mean, and (cf. Exercise 39.12) that its variance is unbiasedly estimated by

$$\hat{V}(t) = \frac{1}{n(n-1)} \sum_{i=1}^r n_i (y_i - t)^2.$$

(Sampford (1962); Pathak (1964b) improves the estimator—cf. 39.5.)

40.4 Show that if k strata of fixed sizes N_i are formed by random subdivision of a population of size N , and then n observations are sampled with equal probabilities without replacement using a USF, the variance of the estimator (39.38) over the entire procedure exactly equals that of the mean of an unstratified sample of the same size from the original population.

Show further that if any allocation but a USF is used in the strata, the variance of (39.38) is increased.

40.5 A sampling design, with one or more stages, selects n first-stage units with replacement from a single stratum of N units, using unequal probabilities $p_r \left(\sum_{i=1}^N p_r = 1 \right)$ at each drawing, and subsequent stages are sampled independently within the selected first-stage units. This design is modified by first dividing the N first-stage units at random into n groups containing N_1, N_2, \dots, N_n units, and selecting one unit from each group with the same relative probabilities as in the original scheme, p_r/p_i , where p_i is the sum of the p_r in the i th group. Subsequent stages remain unchanged. Show that $t_M(z) = \sum_{i=1}^n p_i z_i$ has the same expectation in

the modified design as has $t_0(z) = \frac{1}{n} \sum_{i=1}^n z_i$ in the original design, and that

$$V\{t_M(z)\} = n \frac{\left(\sum_{i=1}^n N_i^2 - N \right)}{N(N-1)} V\{t_0(z)\} + n \frac{\left(N^2 - \sum_{i=1}^n N_i^2 \right)}{N(N-1)}$$

[component of $V\{t_0(z p_i^{\frac{1}{2}})\}$ due to stages after the first].

40.6 In Exercise 40.5, show that if the N_i are chosen to be equal, $t_M(z)$ is never less efficient than $t_0(z)$ for single-stage sampling, while it is given the maximum chance of being no less efficient in multi-stage sampling (cf. Exercise 40.4).

(The results in this and the previous two exercises are given by Stuart (1964), who develops a generalization of the modified sampling scheme; see also J. N. K. Rao *et al.* (1962).)

40.7 In an equal-probabilities random sample of n individuals from a population of N , only n_1 provide the information requested about the variable y . Of the remaining $n - n_1 = n_2$ non-respondents, 1 in every k are subsampled with equal probabilities, giving a subsample of size $r_2 = n_2/k$. The cost function of the sample is

$$C = c_0 n + c_1 n_1 + c_2 n_2.$$

The estimator of μ_y is

$$\hat{\mu}_y = \frac{1}{n}(n_1 m_1 + n_2 m_2),$$

where m_1, m_2 are the means of the respondent sample and the subsample of non-respondents respectively. Show, using (39.72-3), that $\hat{\mu}_y$ is unbiased, with sampling variance

$$V(\hat{\mu}_y) = \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n} + W_2 (k-1) \frac{\sigma_2^2}{n},$$

where σ^2 is the population variance of y and σ_2^2 is the population variance among non-respondents in the population, who form a proportion W_2 of the population. Show, using 39.20, that $V(\hat{\mu}_y)$ is minimized for fixed expected total cost if we choose k to satisfy

$$k_{MV}^2 = \frac{c_2(\sigma^2 - W_2 \sigma_2^2)}{\sigma_2^2 \{c_0 + c_1(1 - W_2)\}}.$$

(cf. Hansen and Hurwitz, 1946)

40.8 In Exercise 40.7, show that if $k = 1$, so that non-respondents are fully sampled, if $\sigma^2 = \sigma_2^2$ and $N \rightarrow \infty$, and a sample is taken with the same expected total cost as the MV sample with $k = k_{MV}$, then its estimation efficiency is given by

$$\frac{V(\hat{\mu}_y | k_{MV})}{V(\hat{\mu}_y | k = 1)} = 1 - \frac{W_2(1 - W_2) \left(1 - \frac{1}{k_{MV}}\right)^2}{W_2 + (1 - W_2)/k_{MV}^2}.$$

Show that if $W_2 = 0.6$ and $k_{MV}^2 \leq 3$, the efficiency with $k = 1 \geq 94$ per cent, illustrating the relative insensitivity of efficiency to departures from k_{MV} .

(Durbin, 1954)

40.9 A very large population is sampled with equal probabilities on two successive occasions. Its variance σ^2 is the same on both occasions, and there is correlation ρ between the values on the two occasions. The first sample is of size n , with mean m_1 . The second sample retains a fraction f of the first sample (with mean m'_1 on the first occasion, m'_2 on the second occasion) and replaces the remaining $n(1-f)$ members of the first sample (with mean m_1^0) by a fresh sample of the same size, with mean m_2'' . Two independent estimators of the population mean on the second occasion are m'_2 and

$$\hat{\mu}_{12} = m'_2 + b_{21}(m_1 - m'_1),$$

a two-phase regression estimator based on the observed regression coefficient b_{21} of the second-occasion values upon the first-occasion values in the nf retained observations. Show that

$$V(\hat{\mu}_{12}) \doteq \frac{\sigma^2}{nf} \{1 - (1-f)\rho^2\},$$

and hence that the MV linear combination of $\hat{\mu}_{12}$ and m_2'' is

$$\hat{\mu} = \frac{f}{\{1 - (1-f)^2 \varrho^2\}} \hat{\mu}_{12} + \frac{(1-f)\{1 - (1-f)\varrho^2\}}{\{1 - (1-f)^2 \varrho^2\}} m_2'',$$

with sampling variance

$$V(\hat{\mu}) \doteq \frac{\sigma^2}{n} \left\{ \frac{1 - (1-f)\varrho^2}{1 - (1-f)^2 \varrho^2} \right\}.$$

40.10 In Exercise 40.9, show further that the change in the population mean between the two occasions may be estimated by

$$d_1 = m_2' - m_1'$$

or

$$d_2 = m_1'' - m_1^0$$

which are independent. Show that the MV linear combination of d_1 and d_2 is

$$d = \frac{f}{\{1 - \varrho(1-f)\}} d_1 + \frac{(1-f)(1-\varrho)}{\{1 - \varrho(1-f)\}} d_2,$$

with sampling variance

$$V(d) \doteq \frac{2\sigma^2}{n} \frac{(1-\varrho)}{\{1 - \varrho(1-f)\}}.$$

(Yates (1960); Patterson (1950) treats the theory for several occasions; Vos (1964) gives variance formulae for simultaneous sampling in time and space.)

40.11 A simple random sample of size n is drawn from an infinite population consisting of k strata containing fractions p_l of the population, the achieved sample in the l th stratum being n_l , $\sum_l n_l = n$. The intended stratum sample sizes are m_l , fixed in advance of sampling,

so a supplementary simple random sample of size $m_l - n_l$ is taken independently within each stratum for which $m_l > n_l$. If the cost per sample member in the initial sample is c , and the cost per sample member in the supplementary sample within the l th stratum is $c_l > c$, while the value of a "surplus" member in the l th stratum (in the initial sample) is $c'_l < c$, show that the expected cost of achieving the intended stratified sample is

$$E(C) = nc + \sum_l [\text{Prob} \{n_l < m_l\} E(m_l - n_l | n_l < m_l) c_l + \text{Prob} \{n_l \geq m_l\} E(m_l - n_l | n_l \geq m_l) c'_l],$$

and that if the n_l are large enough this is approximately

$$E(C) \doteq nc + \sum_l (c_l - c'_l) \left[(m_l - np_l) G \left\{ \frac{m_l - np_l}{[np_l(1-p_l)]^{1/2}} \right\} + [np_l(1-p_l)]^{1/2} g \left\{ \frac{m_l - np_l}{[np_l(1-p_l)]^{1/2}} \right\} \right] + \sum_l (m_l - np_l) c'_l,$$

where G is the standardized normal d.f. and g its f.f. Show that if n is increased by unity, the change in $E(C)$ is

$$\Delta E(c) = c - \sum_l [\text{Prob} \{n_l < m_l\} (c_l - c'_l) + c'_l] p_l$$

and that the least value of n for which $\Delta E(C) > 0$ approximately satisfies

$$\sum_l \left\{ \frac{(c_l - c'_l) p_l}{c - \sum_l c'_l p_l} \right\} \text{Prob} \{n_l < m_l\} = 1,$$

reducing to

$$\sum_l c_l p_l \text{Prob} \{n_l < m_l\} = c$$

when $c'_l = 0$ (i.e. when surplus observations are valueless in any stratum), and to

$$\text{Prob}\{n_l < m_l\} = \frac{c - c'_l}{c_l - c'_l}$$

when all c_l are equal, all c'_l are equal, and all $p_l = \frac{1}{k}$.

(Johnson (1957); Young (1961) considers a related sequential scheme in which the whole population is sampled until every m_l is achieved.)

40.12 Using (40.33), verify the expressions (40.44–9) for the expected values and variances of the three statistics discussed in 40.9.

40.13 Form the differences $V\left(\frac{m_y}{m_x}\right) - V\left\{t\left(\frac{m_y}{m_x}\right)\right\}$ and $V\left\{t\left(\frac{m_y}{m_x}\right)\right\} - V(u)$ in 40.9, and show that these are positive. (Tin, 1965)

40.14 In 40.9, show that the estimator

$$b = \frac{m_y}{m_x} \frac{\left(1 + \alpha_1 \frac{s_{xy}}{m_x m_y}\right)}{\left(1 + \alpha_1 \frac{s_x^2}{m_x^2}\right)}$$

has to order n^{-2} the expected value

$$E(b) = \frac{\mu_y}{\mu_x} \left\{ 1 - \left(2\alpha_2 - \frac{3\alpha_1}{N} \right) (C_{21} - C_{30}) - 2\alpha_1^2 C_{20} (C_{20} - C_{11}) \right\}$$

and the same variance as u defined at (40.32), so that b and u are virtually equivalent.

(Tin (1965)—the estimator is due to E. M. L. Beale.)

40.15 Show that (40.90) may be estimated by

$$\hat{V}(m^{(d)}) = \frac{(N-n)}{N(n^{(d)})^2} \sum_{l=1}^k \frac{n_l}{n_l - 1} \left[\sum_{j=1}^{n_l} (y_{lj}^{(d)} - m_l^{(d)})^2 + n_l^{(d)} \left(1 - \frac{n_l^{(d)}}{n_l} \right) (m_l^{(d)} - m^{(d)})^2 \right],$$

where $m_l^{(d)}$ is the sample analogue of $\mu_l^{(d)}$.

(Durbin, 1958)

40.16 In 40.36, establish (40.94), the generalization of (40.90), and (40.95), generalizing Exercise 40.15.

(Durbin, 1958)

40.17 Verify the numerical values for the mean-square error of the six estimators in (40.29).

CHAPTER 41

MULTIVARIATE DISTRIBUTION THEORY

41.1 In a broad sense nearly the whole of this volume is devoted to multivariate analysis; that is to say, to the analysis of systems in which each member bears the values of more than one measured or classificatory variable. Up to the present, however, we have usually managed to simplify the problems which arise: either (as for sample surveys) being concerned primarily with the estimation of one particular parameter such as a mean, or (as in experimental design) arranging our regressor variables so that estimators of regression coefficients are orthogonal and allow of isolation of individual classification effects. We must now go further and consider systems of greater generality in which the variables are interdependent. In the present chapter we discuss some of the distributional problems which arise. Unless the contrary is stated, the underlying distributions will be assumed to be multivariate normal. It is an unfortunate feature of this branch of the subject that, in other cases, very little is known about exact distribution theory.

41.2 In Chapter 15, Vol. 1, we wrote the p -dimensional multivariate normal distribution in two forms:

$$dF \propto \exp \left\{ -\frac{1}{2} \sum_{j=1}^p \sum_{k=1}^p \alpha_{jk} \left(\frac{x_j - \mu_j}{\sigma_j} \right) \left(\frac{x_k - \mu_k}{\sigma_k} \right) \right\} \prod_{j=1}^p \frac{dx_j}{\sigma_j},$$

$$-\infty \leq x_j \leq \infty, j = 1, 2, \dots, p \quad (41.1)$$

$$dF = \frac{|\alpha|^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}p}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\alpha} (\mathbf{x} - \boldsymbol{\mu}) \right\} \prod_j dx_j, \quad (41.2)$$

where μ_j, σ_j^2 are the mean and variance of the j th variable and $\boldsymbol{\alpha}$ is a matrix inverse to the dispersion matrix.

We were not, in that chapter, much concerned with sampling problems, but we shall now require to distinguish between parent and sample values, or between parameters and estimators. We shall accordingly write a_{jk} for the sample value of α_{jk} , and c_{jk} for the sample covariance whose parent value is γ_{jk} , so that

$$\gamma_{jk} = \rho_{jk} \sigma_j \sigma_k, \quad (41.3)$$

$$c_{jk} = r_{jk} s_j s_k. \quad (41.4)$$

The dispersion matrix which, in Chapter 15, we wrote as \mathbf{V} will now be written $\boldsymbol{\gamma}$, so that we have

$$\boldsymbol{\alpha} = \boldsymbol{\gamma}^{-1}. \quad (41.5)$$

We recall from (15.15) (Vol. 1) that the characteristic function of (41.2) is given by

$$\phi(\mathbf{t}) = \exp \left(-\frac{1}{2} \mathbf{t}' \boldsymbol{\gamma} \mathbf{t} \right) \exp (i \mathbf{t}' \boldsymbol{\mu}). \quad (41.6)$$

41.3 A sample of n values, typified by x_{jl} , $l = 1, 2, \dots, n$, will yield a likelihood function which is the product of n terms of type (41.1), and its logarithm will be the

sum of n terms. Denoting by S summation over sample, and by Σ summation over variables, we find

$$\frac{\partial}{\partial \mu_j} \log L = S \sum_{k=1} \frac{\alpha_{jk}}{\sigma_j} \left(\frac{x_k - \mu_k}{\sigma_k} \right) = 0 \quad (41.7)$$

leading to

$$\sum \frac{\alpha_{jk}}{\sigma_j \sigma_k} (\bar{x}_k - \hat{\mu}_k) = 0. \quad (41.8)$$

Since α is not degenerate, the p equations of this type are equivalent to

$$\hat{\mu}_k = \bar{x}_k, \quad k = 1, 2, \dots, p. \quad (41.9)$$

It is no surprise to find that the sample mean is the ML estimator of the parent mean.

41.4 For the parameter α_{jk} we have

$$\frac{\frac{1}{2}n}{|\alpha|} \frac{\partial |\alpha|}{\partial \alpha_{jk}} - \frac{1}{2} S(x_j - \mu_j)(x_k - \mu_k) = 0. \quad (41.10)$$

If A_{jk} is the co-factor of α_{jk} in $|\alpha|$, we find on substituting for μ the corresponding \bar{x} ,

$$\hat{A}_{jk}/|\alpha| = \frac{1}{n} S(x_j - \mu_j)(x_k - \mu_k) = c_{jk}. \quad (41.11)$$

It follows(*) that

$$\hat{\gamma}_{jk} = c_{jk}. \quad (41.12)$$

In particular, the sample variances are ML estimators of the parent variances, and we also have for the correlations

$$\hat{\rho}_{jk} = r_{jk} = \frac{S(x_j - \bar{x}_j)(x_k - \bar{x}_k)}{\{S(x_j - \bar{x}_j)^2 S(x_k - \bar{x}_k)^2\}^{\frac{1}{2}}}. \quad (41.13)$$

This applies when all the parameters are under estimate. We shall not be concerned with other cases, which are of very minor practical importance, but see Examples 18.14–15 (Vol. 2) and Exercise 18.14 for the bivariate case.

41.5 In setting confidence intervals for these parameters we encounter the same difficulties as in the univariate case, requiring distributions of the “Student” or χ^2 type. We also have a new problem, that of setting simultaneous confidence intervals to the components of a vector. Consider, for example, the estimation of means when parent dispersions are known.

We saw in Chapter 15 that the variables in (41.1) could, by a linear transformation, be reduced to independent normal variables with unit variances. It follows that

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\alpha} (\mathbf{x} - \boldsymbol{\mu})$$

is distributed as χ^2 with p degrees of freedom. We shall show shortly (41.6) that the

(*) This is not, perhaps, immediately obvious. We appeal to a theorem, easily proved (cf. 8.9), that if $\phi_1, \phi_2, \dots, \phi_m$ are functions of parameters $\theta_1, \theta_2, \dots, \theta_m$, then ML estimators of the ϕ 's are obtained by substituting in the functional relations the ML estimators of the θ 's. It also follows that the ML estimators of partial and multiple correlations are given by the corresponding sample statistics.

distribution of means in multivariate normal samples is the same as the distribution of the original variables, except that the dispersion matrix is α/n . It follows, in view of (41.5), that

is distributed as χ^2 with p d.fr. Thus, to a given probability level P we may make assertions of the type

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \boldsymbol{\gamma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \geq \chi_p^2 \quad (41.14)$$

Since we are assuming $\boldsymbol{\gamma}$ known, this sets up a confidence region in the form of a quadric in p dimensions. The practical interpretation of the result requires delicate handling. We shall consider questions of bias in estimation in the next chapter. For the present we are concerned with distributions.

$$\text{Prob } \{n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \boldsymbol{\gamma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \geq \chi_p^2\} = P. \quad (41.15)$$

Wishart's distribution

41.6 We now proceed to investigate the joint distribution of means and dispersions in multivariate normal variation. Suppose we have a random sample of n individuals. Writing x_{jl} for the l th observation on the j th variate, we may array the matrix of observations as

$$\mathbf{x} = (x_{jl}) = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{pmatrix}. \quad (41.16)$$

The frequency distribution of the sample is then given by

$$dF = \frac{|\alpha|^{\frac{1}{2}n}}{(2\pi)^{\frac{1}{2}np}} \exp \left\{ -\frac{1}{2} \sum_{l=1}^n \sum_{j,k=1}^p \alpha_{jk} (x_{jl} - \mu_j) (x_{kl} - \mu_k) \right\} \prod_{l=1}^n \prod_{j=1}^p dx_{jl}. \quad (41.17)$$

We already know from Example 11.7, and 16.25, in Vol. 1 that for $p = 1, 2$, the distribution can be split into two independent distributions, one of means and the other of dispersions. We prove, first of all, that the same is true of any value of p . We have the familiar algebraic identity

$$\sum_{l,j,k} \alpha_{jk} (x_{jl} - \mu_j) (x_{kl} - \mu_k) \equiv S \sum_{j,k} \alpha_{jk} (x_{jl} - \bar{x}_j) (x_{kl} - \bar{x}_k) + n \sum_{j,k} \alpha_{jk} (\bar{x}_j - \mu_j) (\bar{x}_k - \mu_k). \quad (41.18)$$

Thus the exponent in (41.17) factorizes into two components. We now have to consider the differential elements. It will be convenient to make on each variable an orthogonal Helmert transformation of the type used in Example 11.3, Vol. 1:

$$\left. \begin{aligned} y_1 &= \frac{1}{\sqrt{2}} (x_1 - x_2) \\ y_2 &= \frac{1}{\sqrt{6}} (x_1 + x_2 - 2x_3) \\ &\vdots \\ y_{n-1} &= \frac{1}{\sqrt{\{n(n-1)\}}} \{x_1 + x_2 + \dots + x_{n-1} - (n-1)x_n\} \\ y_n &= \frac{1}{\sqrt{n}} (x_1 + x_2 + \dots + x_n) = \sqrt{n} \bar{x}, \end{aligned} \right\} \quad (41.19)$$

where the suffixes are sample labels. The inverse relationship is

$$\left. \begin{aligned} x_1 &= \frac{1}{\sqrt{2}}y_1 + \frac{1}{\sqrt{6}}y_2 + \dots + \frac{1}{\sqrt{\{n(n-1)\}}}y_{n-1} + \frac{1}{\sqrt{n}}y_n \\ x_2 &= -\frac{1}{\sqrt{2}}y_1 + \frac{1}{\sqrt{6}}y_2 + \dots + \frac{1}{\sqrt{\{n(n-1)\}}}y_{n-1} + \frac{1}{\sqrt{n}}y_n \\ &\vdots \\ x_n &= \dots \dots \dots -\frac{n-1}{\sqrt{\{n(n-1)\}}}y_{n-1} + \frac{1}{\sqrt{n}}y_n \end{aligned} \right\}. \quad (41.20)$$

$|J| = 1$ and the differential element in (41.17) becomes simply

$$\prod_{l=1}^n \prod_{j=1}^p dy_{jl}. \quad (41.21)$$

Looking now to (41.18) and remembering that $y_n = \sqrt{n}\bar{x}$ we see that the second factor is a function only of y_{jn} , ($j = 1, 2, \dots, p$). The first factor, in virtue of (41.20), depends only on y_1, \dots, y_{n-1} .

Hence we may factorize off from the original density element the second term in (41.18) and an associated differential element in the \bar{x} 's. With an appropriate adjustment to the constant factor we then obtain for the means

$$dF = \frac{|\alpha|^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}p}} \exp \left\{ -\frac{1}{2}n \sum_{j,k} \alpha_{jk} (\bar{x}_j - \mu_j) (\bar{x}_k - \mu_k) \right\} \prod_{j=1}^p d(\sqrt{n}\bar{x}_j). \quad (41.22)$$

The joint distribution of means is, in fact, the same as that of the original variables, apart from the factor in n .

41.7 The distribution of the sample variances and covariances is thus confined to the $(n-1)$ -spaces orthogonal to the sample means. Since the orthogonal transformation is simply a rotation of axes, it leaves distances and angles invariant. Since variances and covariances are functions of these alone (cf. Example 11.7 and 16.24), they too are invariant. The non-differential part arising from (41.18) and (41.22) may be written

$$f = \frac{|\alpha|^{\frac{1}{2}(n-1)}}{(2\pi)^{\frac{1}{2}(n-1)p}} \exp \left(-\frac{1}{2}n \sum_{j,k} \alpha_{jk} c_{jk} \right). \quad (41.23)$$

Our principal problem is to evaluate the differential element in terms of the c_{jk} . Write

$$u_{jl} = y_{jl} - \bar{y}_j. \quad (41.24)$$

Then the covariance c_{jk} is given by

$$nc_{jk} = \sum_{l=1}^n u_{jl} u_{kl}. \quad (41.25)$$

Note that here and throughout we use n as the sample number, not the degree of freedom of the dispersions. We require that

$$n-1 \geq p.$$

Generalizing the argument of 16.24, Vol. 1, we take p flat spaces of $n-1$ dimensions, one for each u , and let the sample points be represented by P_1, P_2, \dots, P_p . We shall consider in turn the variation in P_1 , then that of P_2 for fixed P_1 , then that of P_3 for fixed P_1 and P_2 , and so on. We shall then multiply all these together to find the variation of P_1, P_2, \dots, P_p .

Consider then P_m given P_1, P_2, \dots, P_{m-1} . Let O be the origin and imagine the spaces superposed on one another. For fixed OP_m and fixed angles $P_m OP_1, P_m OP_2, \dots, P_m OP_{m-1}$, the point P_m varies on a hypersphere of $n-m$ dimensions. Let the length of the perpendicular line from P_m on to the hyperplane determined by O, P_1, \dots, P_{m-1} be t_m . Then the "content" (volume) of variation permissible to P_m is that of the "surface" of a hypersphere in $n-m$ dimensions with radius t_m , which is^(*)

$$= \frac{2\pi^{\frac{1}{2}(n-m)} t_m^{n-m-1}}{\Gamma\{\frac{1}{2}(n-m)\}}. \quad (41.26)$$

We require to multiply this by the element of variation perpendicular to the hypersphere. Consider the transformation in the m -space determined by O, P_1, \dots, P_m , based on (41.25),

$$\xi_{mj} = nc_{mj} = \sum_{k=1}^n u_{mk} u_{jk}, \quad j = 1, 2, \dots, m. \quad (41.27)$$

The Jacobian of the transformation is given by

$$J = \frac{\partial(\xi_{m1}, \dots, \xi_{mm})}{\partial(u_{m1}, \dots, u_{mm})} = \begin{vmatrix} u_{11} & u_{12} & \dots & u_{1m} \\ u_{21} & u_{22} & \dots & u_{2m} \\ \dots & \dots & \dots & \dots \\ 2u_{m1} & 2u_{m2} & \dots & 2u_{mm} \end{vmatrix}. \quad (41.28)$$

and this is equal to $2v_m$ where v_m is the volume of the parallelotope (the m -dimensional parallelogram) determined by O, P_1, \dots, P_m . Thus the differential element is

$$\frac{1}{2v_m} \prod_{j=1}^m d\xi_{mj}. \quad (41.29)$$

On multiplication by (41.26) the total variation of P_m is then given by

$$\frac{\pi^{\frac{1}{2}(n-m)} t_m^{n-m-1}}{\Gamma\{\frac{1}{2}(n-m)\}} \prod_{j=1}^m d\xi_{mj}. \quad (41.30)$$

But we have

$$t_m = \frac{v_m}{v_{m-1}}, \quad (41.31)$$

$$\text{and, from (41.27),} \quad |\xi_{mj}| = |u_{mj}|^2 = v_m^2. \quad (41.32)$$

The element of variation (41.30) then becomes

$$\frac{\pi^{\frac{1}{2}(n-m)} v_m^{n-m-2}}{\Gamma\{\frac{1}{2}(n-m)\} v_{m-1}^{n-m-1}} \prod d\xi_{mj}. \quad (41.33)$$

We now multiply expressions of type (41.33) for $m = 1, 2, \dots, p$. The terms in v cancel except for those in v_0 (which is unity) and v_p , and we find

$$\frac{\pi^{\frac{1}{2}p(2n-p-1)}}{\prod_{j=1}^p \Gamma\{\frac{1}{2}(n-j)\}} v_p^{n-p-2} \prod_{j=1}^p \prod_{k=m}^p d\xi_{jk}. \quad (41.34)$$

$$v_p^2 = |\xi_{pj}| = n^p |c_{jk}|. \quad (41.35)$$

From (41.32) we have

(*) Cf. Kendall, M. G., *A Course in the Geometry of n Dimensions*, p. 42.

Putting (41.34-5) and (41.23) together, we then have for the distribution of dispersions

$$dF = \frac{(\frac{1}{2}n)^{\frac{1}{2}p(n-1)} |\alpha|^{\frac{1}{2}(n-1)} |c|^{\frac{1}{2}(n-p-2)}}{\pi^{\frac{1}{2}p(p-1)} \prod_{j=1}^p \Gamma\{\frac{1}{2}(n-j)\}} \exp(-\frac{1}{2}n \sum \alpha_{jk} c_{jk}) \prod_{j < k}^p dc_{jk}. \quad (41.36)$$

This is Wishart's (1928) distribution. (11.41) in Example 11.7 is the case $p = 1$; (16.54) is a simple transformation of the case $p = 2$.

Readers who find the geometrical line of proof hard to follow may prefer to consult the review of alternative proofs in Wishart's article (1948). From many points of view the distribution may be regarded as the generalization of the χ^2 distribution to p -dimensional variation.

41.8 Let us note some minor but not unimportant details concerning the distribution:

- (a) In the exponent of (41.36) we are summing over all j, k and, since $\alpha_{jk} = \alpha_{kj}$, $c_{jk} = c_{kj}$, the terms occur in pairs except when $j = k$. Thus there are p terms of type $\alpha_{11}c_{11}$ and $\frac{1}{2}p(p-1)$ of type $\alpha_{12}c_{12}$. For example, with $p = 2$ the exponent is $-\frac{1}{2}n(\alpha_{11}c_{11} + 2\alpha_{12}c_{12} + \alpha_{22}c_{22})$.
- (b) In the differential element there are $\frac{1}{2}p(p+1)$ terms, not p^2 . Thus, for $p = 2$ the element is $dc_{11}dc_{12}dc_{22}$.
- (c) The domain of variation for each c_{jj} is 0 to ∞ , but we cannot easily specify those for the other c 's, which are conditioned by the fact that the matrix (c_{jk}) must be positive definite. This makes it extremely difficult to integrate out certain variables c in order to obtain the marginal distributions of others.
- (d) We have defined the sample variances and covariances by dividing the appropriate product-sum by n . We may, if we prefer, divide by $n-1$, in which case appropriate adjustments have to be made in (41.36). The reader should watch this point in consulting the literature, because usage varies.

41.9 We can now derive the characteristic function of the Wishart distribution. Writing a single integral sign to denote integration over the domain of c 's, we have from (41.36)

$$\int |c|^{\frac{1}{2}(n-p-2)} \exp(-\frac{1}{2}n \sum \alpha_{jk} c_{jk}) \prod dc_{jk} = k |\alpha|^{-\frac{1}{2}(n-1)} \quad (41.37)$$

where k is some constant. If we replace

α_{jj} by $\alpha_{jj} - 2\theta_{jj}/n$, α_{jk} by $\alpha_{jk} - \theta_{jk}/n$, $j \neq k$,

the exponent under the integral sign gives us the c.f. of the c 's with θ_{jk} for the usual imaginary dummy variable it_{jk} . Making the substitutions on the right in (41.37), and adjusting the constant to make ϕ unity for zero θ , we have

$$\phi(\theta) = \frac{|\alpha|^{\frac{1}{2}(n-1)}}{\begin{vmatrix} \alpha_{11} - 2\theta_{11}/n & \alpha_{12} - \theta_{12}/n & \dots & \alpha_{1p} - \theta_{1p}/n \\ \alpha_{21} - \theta_{21}/n & \alpha_{22} - 2\theta_{22}/n & \dots & \alpha_{2p} - \theta_{2p}/n \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{p1} - \theta_{p1}/n & \alpha_{p2} - \theta_{p2}/n & \dots & \alpha_{pp} - 2\theta_{pp}/n \end{vmatrix}}^{\frac{1}{2}(n-1)} \quad (41.38)$$

This substitutional device, avoiding the problem of actually integrating over the domain of variation, is a useful one which we shall employ again.

Example 41.1

In the bivariate case ($p = 2$) we have, in the usual notation (with unit variances),

$$\gamma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad |\gamma| = 1 - \rho^2,$$

$$\alpha = \begin{vmatrix} \frac{1}{1-\rho^2} & -\frac{\rho}{1-\rho^2} \\ -\frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{vmatrix}, \quad |\alpha| = (1-\rho^2)^{-1}.$$

Thus (41.36) reduces to

$$dF = \frac{(\frac{1}{2}n)^{n-1}}{(1-\rho^2)^{\frac{1}{2}(n-1)}} \frac{s_1^{n-4} s_2^{n-4} (1-\rho^2)^{\frac{1}{2}(n-4)}}{\pi^{\frac{1}{2}} \Gamma\{\frac{1}{2}(n-1)\} \Gamma\{\frac{1}{2}(n-2)\}} \times \exp \left\{ \frac{-n}{2(1-\rho^2)} (s_1^2 - 2\rho r s_1 s_2 + s_2^2) \right\} d(s_1^2) d(s_2^2) d(rs_1 s_2).$$

On using the duplication formula

$$\Gamma\{\frac{1}{2}(n-1)\} \Gamma\{\frac{1}{2}(n-2)\} = \frac{\pi^{\frac{1}{2}} \Gamma(n-2)}{2^{n-3}}$$

we find

$$dF = \frac{n^{n-1}}{(1-\rho^2)^{\frac{1}{2}(n-1)}} \frac{s_1^{n-4} s_2^{n-4} (1-\rho^2)^{\frac{1}{2}(n-4)}}{4\pi \Gamma(n-2)} \times \exp \left\{ \frac{-n}{2(1-\rho^2)} (s_1^2 - 2\rho r s_1 s_2 + s_2^2) \right\} d(s_1^2) d(s_2^2) d(rs_1 s_2) \quad (41.39)$$

which reduces to the form found in 16.26.

Example 41.2

Let us now consider the moments of the distribution of the covariance when $p = 2$. From (41.38) we have, putting $\theta_{11} = \theta_{22} = 0$,

$$\phi(\theta_{12}) \propto \begin{vmatrix} \frac{1}{1-\rho^2} & -\frac{\rho}{1-\rho^2} - \frac{\theta_{12}}{n} \\ -\frac{\rho}{1-\rho^2} - \frac{\theta_{12}}{n} & \frac{1}{1-\rho^2} \end{vmatrix}^{-\frac{1}{2}(n-1)}. \quad (41.40)$$

Expanding and evaluating the constant from the consideration that $\phi(0) = 1$ we find

$$\phi(\theta_{12}) = \left\{ 1 - \frac{2\rho\theta_{12}}{n} - \frac{(1-\rho^2)\theta_{12}^2}{n^2} \right\}^{-\frac{1}{2}(n-1)}. \quad (41.41)$$

Taking logarithms and evaluating coefficients of θ_{12} we find

$$\kappa_1 = \frac{n-1}{n} \rho \quad (41.42)$$

$$\kappa_2 = \frac{n-1}{n^2} (1+\rho^2) \quad (41.43)$$

$$\kappa_3 = \frac{2(n-1)}{n^3} \rho(3+\rho^2) \quad (41.44)$$

$$\kappa_4 = \frac{6(n-1)}{n^4} (1+6\rho^2+\rho^4). \quad (41.45)$$

In standard measure the distribution tends to normality as n tends to infinity. But for finite n

$$\beta_1^2 = \frac{4}{n-1} \frac{\rho^2(3+\rho^2)^2}{(1+\rho^2)^3} \quad (41.46)$$

$$\beta_2 = 3 + \frac{6}{n-1} \frac{1+6\rho^2+\rho^4}{(1+\rho^2)^2}. \quad (41.47)$$

Thus, even for $\rho = 0$, the distribution, though symmetrical, is not normal. We have derived these results differently in Example 13.3, Vol. 1.

Wishart (1929) gave the formulae explicitly for $p \leq 8$ as far as those of the fourth order.

The additive property of Wishart distributions

41.10 One property of the Wishart distribution, analogous to that of χ^2 in one dimension, is worth noticing.

Suppose we have two samples, n_1 and n_2 in number, from the same multivariate distribution. If we pool them, of course, the dispersions of the total sample will follow the Wishart distribution for a sample of $n_1 + n_2$. But we may also consider the joint distribution of the dispersions from each sample. If we form a new dispersion matrix by adding corresponding dispersions, i.e.

$$c_{jk} = c_{jk}^{(1)} + c_{jk}^{(2)} \quad (41.48)$$

where the superfixes refer to the first and second sample, then the c_{jk} 's are distributed in Wishart's form with $n_1 + n_2 - 1$ instead of n .

This is perhaps most easily seen from the characteristic function. If we adjoin the distributions of $c^{(1)}$ and $c^{(2)}$ it will be clear, as in (41.37), that the c.f. of c itself is of the same form as (41.38).

41.11 It would add a pleasant completeness to the sampling theory of dispersions if we could proceed from the Wishart distribution and deduce the distribution of particular functions of the variances and covariances by integrating over appropriate domains to eliminate unwanted variables. Unfortunately this is prohibitively complicated in general. As in Example 41.2 we can derive moments and product-moments of sample dispersions; finding the explicit distributions is then usually unnecessary. There are, however, a few cases in which we can proceed further.

Moments of the dispersion determinant (generalized variance)

41.12 Consider the distribution of the dispersion determinant $|c|$. From (41.36), the integral being taken over the variances and covariances,

$$\int \frac{(\frac{1}{2}n)^{\frac{1}{2}p(n-1)} |\alpha|^{\frac{1}{2}(n-1)} |c|^{\frac{1}{2}(n-p-2)}}{n^{\frac{1}{2}p(p-1)} \prod_{j=1}^p \Gamma\{\frac{1}{2}(n-j)\}} \exp(-\frac{1}{2}n \sum \alpha_{jk} c_{jk}) \prod_{j \leq k}^p dc_{jk} = 1. \quad (41.49)$$

Put $n\alpha_{jk} = \beta_{jk}$. We have then

$$\int \frac{|\beta|^{\frac{1}{2}(n-1)} |c|^{\frac{1}{2}(n-p-2)}}{\pi^{\frac{1}{2}p(p-1)}} \exp\left(-\frac{1}{2} \sum \beta_{jk} c_{jk}\right) \Pi dc_{jk} = 2^{\frac{1}{2}p(n-1)} \prod_{j=1}^p \Gamma\left\{\frac{1}{2}(n-j)\right\}. \quad (41.50)$$

Replace n by $n+2t$ and divide by $|\beta|^t$. We then have

$$\int \frac{|\beta|^{\frac{1}{2}(n-1)} |c|^{\frac{1}{2}(n-p-2)+t}}{\pi^{\frac{1}{2}p(p-1)}} \exp\left(-\frac{1}{2} \sum \beta_{jk} c_{jk}\right) \Pi dc_{jk} = \frac{2^{\frac{1}{2}p(n-1)+pt}}{|\beta|^t} \prod_{j=1}^p \Gamma\left\{\frac{1}{2}(n-j)+t\right\}. \quad (41.51)$$

If we now replace β_{jk} by $n\alpha_{jk}$ and divide by $2^{\frac{1}{2}p(n-1)} \prod \Gamma\left\{\frac{1}{2}(n-j)\right\}$ we have on the left the expectation of $|c|^t$. Thus

$$E(|c|^t) = \frac{2^{pt}}{n^{pt} |\alpha|^t} \frac{\prod_{j=1}^p \Gamma\left\{\frac{1}{2}(n-j)+t\right\}}{\prod_{j=1}^p \Gamma\left\{\frac{1}{2}(n-j)\right\}} \quad (41.52)$$

$$= \frac{2^{pt}}{n^{pt}} \prod_{j=1}^p \frac{\Gamma\left\{\frac{1}{2}(n-j)+t\right\}}{\Gamma\left\{\frac{1}{2}(n-j)\right\}} |\gamma|^t. \quad (41.53)$$

From this we can determine as many moments or cumulants as we wish. Again the substitutional device, obviating the awkward integrations, is to be noted.

41.13 One consequence of (41.53) is noteworthy. If we write d_{jk} for the sums of products about the mean, so that $d_{jk} = nc_{jk}$, we have

$$E\left(\frac{|d|^t}{|\gamma|^t}\right) = 2^{pt} \prod_{j=1}^p \frac{\Gamma\left\{\frac{1}{2}(n-j)+t\right\}}{\Gamma\left\{\frac{1}{2}(n-j)\right\}}. \quad (41.54)$$

Now a χ^2 variable with ν degrees of freedom has moments about zero given by

$$\mu'_t = 2^t \frac{\Gamma\left(\frac{1}{2}\nu+t\right)}{\Gamma\left(\frac{1}{2}\nu\right)}.$$

The right-hand side of (41.54) is the product of p such factors with $\nu = n-1, n-2, \dots, n-p$. Remembering that the moment of a product of independent variables is the product of their moments, we see that $|d|/|\gamma|$ can be represented as the product of p independent factors, distributed as χ^2 with $n-1, n-2, \dots, n-p$ degrees of freedom.

Example 41.3

When $p = 1$ we have the familiar result that a sum of squares, standardized by division by a parent variance, is distributed as χ^2 with $n-1$ d.f.

For $p = 2$ we find from (41.54) for the moments of $|d|/|\gamma|$

$$\mu'_t = \frac{\Gamma\left\{\frac{1}{2}(n-1)+t\right\}}{\Gamma\left\{\frac{1}{2}(n-1)\right\}} \frac{\Gamma\left\{\frac{1}{2}(n-2)+t\right\}}{\Gamma\left\{\frac{1}{2}(n-2)\right\}} 2^{2t}. \quad (41.55)$$

From the duplication formula for the Gamma function,

$$\Gamma(x) \Gamma\left(x+\frac{1}{2}\right) = \frac{\pi^{\frac{1}{2}} \Gamma(2x)}{2^{2x-1}}, \quad (41.56)$$

we reduce this to

$$\mu'_t = \frac{\Gamma(n-2+2t)}{\Gamma(n-2)}. \quad (41.57)$$

This is the $(2t)$ th moment of a Gamma variable with parameter $(n-2)$ —cf. Exercise 4.6, Vol. 1. Thus the t th moment of $2|d|/\gamma$ is equal to the $(2t)$ th moment of a χ^2 variable with $2(n-2)$ d.fr. This is a special case of the result of Exercise 11.9, Vol. 1.

The correlation determinant

41.14 There is considerable interest in a study of the joint distribution of sample correlations. In the general case (non-vanishing parent correlations) distributions are complicated, even for $p = 2$ —cf. 16.26–7. We can make some progress in the null case, i.e. when all parental values are zero. The Wishart distribution (41.36) then reduces to

$$dF = \frac{(\frac{1}{2}n)^{\frac{1}{2}p(n-1)} |c|^{\frac{1}{2}(n-p-2)}}{\pi^{\frac{1}{2}p(p-1)} \prod_{j=1}^p \Gamma\{\frac{1}{2}(n-j)\}} \exp \left\{ -\frac{1}{2}n \sum_{j=1}^p c_{jj} \right\} \prod_{j < k} dc_{jk}. \quad (41.58)$$

Transforming to new variables by equations of the type

$$c_{jk} = s_j s_k r_{jk}, \quad (41.59)$$

we find that the Jacobian is given by

$$J = 2^p \prod_{j=1}^p s_j^p. \quad (41.60)$$

This is independent of the r_{jk} , as is the exponent in (41.58), and consequently terms in s can be factorized off, leaving us with the distribution of correlations

$$dF = \frac{|r|^{\frac{1}{2}(n-p-2)} [\Gamma\{\frac{1}{2}(n-1)\}]^p}{\pi^{\frac{1}{2}p(p-1)} \prod_{j=1}^p \Gamma\{\frac{1}{2}(n-j)\}} \prod_{j < k} dr_{jk}, \quad (41.61)$$

where $|r|$ is the sample correlation determinant and the constant has been adjusted so that the frequency function integrates to unity. We are again hindered from making progress by the boundaries of the domain of variation. In the manner of 41.12, however, we can find the moments of $|r|$ itself. They are

$$E|r|^t = \frac{[\Gamma\{\frac{1}{2}(n-1)\}]^p}{[\Gamma\{\frac{1}{2}(n-1)+t\}]^p} \frac{\prod_{j=1}^p \Gamma\{\frac{1}{2}(n-j)+t\}}{\prod_{j=1}^p \Gamma\{\frac{1}{2}(n-j)\}}. \quad (41.62)$$

Example 41.4

Writing $L = \log |r|$, $\nu = \frac{1}{2}(n-1)$, $m = \frac{1}{2}(j-1)$ we have from (41.62)

$$E(e^{Lt}) = \prod_{j=1}^p \frac{\Gamma(\nu)}{\Gamma(\nu+t)} \frac{\Gamma(\nu-m+t)}{\Gamma(\nu-m)}. \quad (41.63)$$

This was shown to be true for integral t only, but by analytic continuation may be shown to be generally true for all t . We may therefore interpret t as an imaginary number, and (41.63) then gives us the characteristic function of L . For the corresponding cumulant-generating function we then have

$$\psi = \sum_{j=1}^p \{\log \Gamma(\nu) + \log \Gamma(\nu-m+t) - \log \Gamma(\nu+t) - \log \Gamma(\nu-m)\}. \quad (41.64)$$

We recall the expansion

$$\log \Gamma(x) = (x - \frac{1}{2}) \log x - x + \frac{1}{2} \log(2\pi) + \frac{1}{12x} - \frac{1}{360x^3} + \dots \quad (41.65)$$

For large ν , substitution in (41.64) gives us for the term in braces

$$-\frac{mt}{\nu} + \frac{1}{\nu^2} \left\{ -\frac{1}{2}m(m+1)t + \frac{1}{2}mt^2 \right\} + O(\nu^{-3}). \quad (41.66)$$

Now put $m = \frac{1}{2}(j-1)$ and sum from $j = 1$ to p . We find

$$\psi = -\frac{1}{4}p(p-1)\frac{t}{\nu} - O\left(\frac{t^2}{\nu^2}\right) + \frac{1}{4}p(p-1)\frac{t^2}{2\nu^2} + O(\nu^{-3}). \quad (41.67)$$

Thus

$$E(L) = -\frac{1}{4}p(p-1)/\nu + O(\nu^{-2}), \quad (41.68)$$

$$\text{var } L = \frac{1}{4}p(p-1)/\nu^2 + O(\nu^{-3}). \quad (41.69)$$

Furthermore we have

$$\left(\frac{d}{dx}\right)^k \log \Gamma(x) = \frac{(-1)^{k-2}(k-2)!}{x^{k-1}} + o(x^{-k+1}), \quad k > 1. \quad (41.70)$$

Hence, from (41.64), to the greatest power in ν ,

$$\begin{aligned} \left[\frac{d^k \psi}{dt^k}\right]_{t=0} &= \sum_{j=1}^p (-1)^{k-2}(k-2)! \left\{ \frac{1}{(\nu-m+t)^{k-1}} - \frac{1}{(\nu+t)^{k-1}} \right\}_{t=0} \\ &= \sum_{j=1}^p \frac{(-1)^{k-2}(k-1)!}{\nu^k} m, \quad \text{where } m = \frac{1}{2}(j-1) \\ &= (-2\nu)^{-k} 2^{k-1}(k-1)! \frac{1}{2}p(p-1). \end{aligned} \quad (41.71)$$

Comparison with equation (16.4), Volume 1, shows that, with a suitable choice of origin, $-2\nu \log |r|$ is asymptotically distributed as χ^2 with $\frac{1}{2}p(p-1)$ degrees of freedom. To order ν^{-1} the origin, from (41.68), is seen to be zero.

Thus $-(n-1) \log |r|$ is asymptotically distributed as χ^2 with $\frac{1}{2}p(p-1)$ d.f. Bartlett (1951) gave a slightly more refined result, namely that

$$-\{n - \frac{1}{6}(2p+1)\} \log |r| \text{ is } \chi^2 \text{ with } \frac{1}{2}p(p-1) \text{ d.f.} \quad (41.72)$$

The extra term derives from an allowance for the item of order n^{-1} in the mean, but in practice this is a refinement of minor importance.

Example 41.5

It is interesting to compare the results of 41.13, concerning the dispersion determinant, with those of the previous example for the correlation determinant in the null case.

Without loss of generality, suppose that the parent variances are unity and the parent correlations zero. Then, from 41.13,

$$n^p |c| = (ns_1^2)(ns_2^2) \dots (ns_p^2) |r|$$

is the product of p independent χ^2 variables with $n-1, n-2, \dots, n-p$ degrees

of freedom, whereas $-(n-1) \log |r|$ is asymptotically χ^2 with $\frac{1}{2}p(p-1)$ degrees of freedom.

Now from (27.61) applied to the sample instead of the population, we have in our present notation

$$1 - R_{1(2 \dots p)}^2 = \frac{|r|}{R_{11}} \quad (41.73)$$

where $R_{1(2 \dots p)}$ is the multiple correlation coefficient of x_1 on x_2, x_3, \dots, x_p and R_{11} is the cofactor of r_{11} ($= 1$) in the correlation determinant. By repeated applications of this formula we have, on re-ordering suffixes,

$$\{1 - R_{p(12 \dots (p-1))}^2\} \{1 - R_{p-1(12 \dots (p-2))}^2\} \dots \{1 - R_{3(12)}^2\} \{1 - R_{2(1)}^2\} = |r| \quad (41.74)$$

where $R_{2(1)}^2$ is the same as the zero order correlation r_{21}^2 . Moreover, all the x 's are independent, so all the factors on the left in (41.74) are independent (cf. 27.30).

The distribution of $U = 1 - R^2$, based on a sample of n and q variables, is (from (27.74))

$$dF = \frac{(1-U)^{\frac{1}{2}(q-3)} U^{\frac{1}{2}(n-q-2)}}{B\{\frac{1}{2}(q-1), \frac{1}{2}(n-q)\}} dU. \quad (41.75)$$

By the same kind of argument as we used in the previous example we can find the moments of U and hence the characteristic function of $\log U$. And in fact $(n-1) \log U$ is found to be distributed approximately as χ^2 with $q-1$ degrees of freedom.

Thus $-(n-1) \log |r|$ is approximately the sum of p independent χ^2 factors with $p-1, p-2, \dots, 1$ degrees of freedom, namely as a χ^2 with $\frac{1}{2}p(p-1)$ d.fr. This checks with the result we obtained in the previous example.

The ratio $|r|/R_{11}$ is also equal (cf. (27.34)) to $s_{1.2 \dots p}^2/s_1^2$. In R_{11} , the corresponding ratio is equal to $s_{2.3 \dots p}^2/s_2^2$, and so on. Thus

$$n^p |c| = \{ns_{1.2 \dots p}^2\} \{ns_{2.3 \dots p}^2\} \dots \{ns_p^2\}. \quad (41.76)$$

The sums of squares of type $ns_{j(k)}^2$ are residuals which are all independent, and are based on $n-p, n-p+1, \dots, n-1$ degrees of freedom. Thus $n^p |c|$ is the product of independent χ^2 factors with those d.fr., confirming the results of 41.13.

Hotelling's T^2

41.15 We proceed to derive a generalization, due to Hotelling, of "Student's" t . As in 41.13, we write $d_{jk} = nc_{jk}$. Let (D_{jk}) be the inverse of (d_{jk}) . Define

$$T^2 = n(n-1) \sum_{j,k=1}^p D_{jk} \bar{x}_j \bar{x}_k. \quad (41.77)$$

When $p = 1$ we have

$$d_{11} = ns_1^2, \quad D_{11} = 1/(ns_1^2) \\ T^2 = \frac{n(n-1) \bar{x}^2}{ns_1^2} = \frac{(n-1) \bar{x}^2}{s_1^2} = t^2,$$

which illustrates how T reduces to "Student's" t in the univariate case.

Let m_{jk} denote the sums of squares about the origin so that

$$m_{jk} = d_{jk} + n\bar{x}_j \bar{x}_k. \quad (41.78)$$

The determinant $|m_{jk}|$ may be written

$$\begin{vmatrix} 1 & \bar{x}_1\sqrt{n} & \bar{x}_2\sqrt{n} & \cdots & \bar{x}_p\sqrt{n} \\ 0 & d_{11} + n\bar{x}_1^2 & d_{12} + n\bar{x}_1\bar{x}_2 & \cdots & d_{1p} + n\bar{x}_1\bar{x}_p \\ 0 & d_{21} + n\bar{x}_2\bar{x}_1 & d_{22} + n\bar{x}_2^2 & \cdots & d_{2p} + n\bar{x}_2\bar{x}_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & d_{p1} + n\bar{x}_p\bar{x}_1 & d_{p2} + n\bar{x}_p\bar{x}_2 & \cdots & d_{pp} + n\bar{x}_p^2 \end{vmatrix}. \quad (41.79)$$

Subtracting $\bar{x}_1\sqrt{n}$ times the first row from the second, $\bar{x}_2\sqrt{n}$ times the first row from the third, and so on, we find

$$|m_{jk}| = \begin{vmatrix} 1 & \bar{x}_1\sqrt{n} & \cdots & \bar{x}_p\sqrt{n} \\ -\bar{x}_1\sqrt{n} & d_{11} & \cdots & d_{1p} \\ -\bar{x}_2\sqrt{n} & d_{21} & \cdots & d_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ -\bar{x}_p\sqrt{n} & d_{p1} & \cdots & d_{pp} \end{vmatrix}. \quad (41.80)$$

Expand by the border row and column. We find

$$|m_{jk}| = |d_{jk}| + \sum_{j,k=1}^p nD_{jk}\bar{x}_j\bar{x}_k |d_{jk}|. \quad (41.81)$$

From (41.77) it then follows that

$$\frac{|d_{jk}|}{|m_{jk}|} = \frac{1}{1 + T^2/(n-1)}. \quad (41.82)$$

41.16 Consider the geometrical interpretation of this result. In the case $p = 1$ the numerator and denominator of (41.82) reduce to d_{11} and m_{11} , that is to say, to the squares of the distances from the sample point P_1 (in the n -dimensional sample space) to its projection on the unit vector whose direction cosines are all equal, and from P_1 to the origin O respectively. The ratio is the square of the sine of the angle between OP_1 and the unit vector. This was the geometrical approach which gave us "Student's" distribution in Example 11.8 (Vol. 1).

In the general case, consider the p superposed sample spaces discussed in the derivation of Wishart's distribution in 41.7. From a relation similar to that of (41.32) we see that $|m_{jk}|$ is the square of the volume of a parallelotope (generalized parallelogram) with one corner at the origin and sides parallel to OP_1, OP_2, \dots, OP_p .

Further, if H is a hyperplane perpendicular to the unit vector meeting it in O' , it is easy to see that the projections of P_1, P_2, \dots, P_p on to H , say P'_1, P'_2, \dots, P'_p , are such that d_{jk} represents sums of squares and cross-products in H referred to O' as origin. Thus $|d_{jk}|$ is the square of the volume of a parallelotope in H . Thus the ratio of $|d_{jk}|$ to $|m_{jk}|$ is the square of the cosine of the angle between H and the unit vector. If the angle is θ we then have

$$\frac{1}{1 + T^2/(n-1)} = \cos^2 \theta. \quad (41.83)$$

Now if the sample points P are distributed at random in the n -spaces, the hyperplane which they determine is distributed at random in regard to the angle which it

makes with a fixed vector, and in particular the unit vector. The sampling distribution of θ is then that of an angle between a fixed vector and a random hyperplane. But this is the same as the distribution of the angle between a fixed hyperplane and a random vector. And this, from a slightly different viewpoint, is the problem of distribution which we solved in connection with the multiple correlation coefficient R , for we saw (27.26, Vol. 2) that R can be regarded as the sine of the angle between a residual variable represented by $x_{1.2 \dots p}$ and the space containing the other variables x_2, \dots, x_p , and when the former is independent of the latter we can regard one of them as fixed. Thus we may write

$$\frac{1}{1 + T^2/(n-1)} = 1 - R^2, \quad (41.84)$$

where we must remember that in the distribution of R^2 , namely

$$dF = \frac{1}{B\{\frac{1}{2}(n-p), \frac{1}{2}(p-1)\}} (1 - R^2)^{\frac{1}{2}(n-p-2)} (R^2)^{\frac{1}{2}(p-3)} dR^2, \quad (41.85)$$

p is the total number of variables and the variate values are measured from their means in forming the regression equation. In applying (41.85) to Hotelling's distribution we must increase p by unity, for we are effectively considering $p+1$ variables—the unit vector being the extra one; and we must increase n by unity because our variation is not restricted to that about the mean. Making these adjustments and substituting in (41.85) from (41.84), we find for the distribution of T^2

$$dF = \frac{1}{B\{\frac{1}{2}(n-p), \frac{1}{2}p\}} \frac{\{T^2/(n-1)\}^{\frac{1}{2}(p-2)}}{\{1 + T^2/(n-1)\}^{\frac{1}{2}n}} d\left(\frac{T^2}{n-1}\right). \quad (41.86)$$

Equivalently, we may say that

$$\frac{(n-p)T^2}{p(n-1)} \text{ has an } F(p, n-p) \text{ distribution.} \quad (41.87)$$

This may be used in the obvious way to test a hypothetical vector of means μ_0 , by measuring \mathbf{x} from this as origin and then using the distribution (41.87).

41.17 The same result may be derived by using the substitutional device of 41.12. Starting from the Wishart distribution, we note that if product-moments about the origin, say c' , are used instead of those about means, the distribution remains valid except that there is an extra degree of freedom. We then have

$$dF = \frac{(\frac{1}{2}n)^{\frac{1}{2}pn} |\alpha|^{\frac{1}{2}n} |c'|^{\frac{1}{2}(n-p-1)}}{\pi^{\frac{1}{2}p(p-1)} \prod \Gamma\{\frac{1}{2}(n+1-j)\}} \exp \left\{ -\frac{1}{2}n \sum \alpha_{jk} c'_{jk} \right\} \prod dc'_{jk}. \quad (41.88)$$

As in (41.53) we find

$$E(|c'|)^t = \frac{2^{pt}}{|\beta|^t} \prod_{j=1}^p \frac{\Gamma\{\frac{1}{2}(n+1-j) + t\}}{\Gamma\{\frac{1}{2}(n+1-j)\}}. \quad (41.89)$$

Now we may also write (41.88) in the form given by our original derivation of Wishart's distribution

$$dF = \frac{|\frac{1}{2}\beta|^{\frac{1}{2}(n-1)} |c|^{\frac{1}{2}(n-p-2)}}{\pi^{\frac{1}{2}p(p-1)} \prod \Gamma\{\frac{1}{2}(n-j)\}} \exp(-\frac{1}{2} \sum \beta_{jk} c_{jk}) \prod dc_{jk}$$

If we multiply this by $|c'|^t$ and integrate we obtain the form on the right in (41.89). Replace n by $n+2u$ and divide by $|\frac{1}{2}\beta|^u$. We then have, on division by appropriate constants,

$$\times \frac{|\beta|^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}p}} \exp(-\frac{1}{2} \sum \beta_{jk} \bar{x}_j \bar{x}_k) \prod d\bar{x}. \quad (41.90)$$

We recall that $|c|/|c'| = |d|/|m|$. Put $t = -u$ in (41.91). We then have

$$E\{|c'|^t |c|^u\} = \frac{2^{p(t+u)}}{|\beta|^{t+u}} \prod_{j=1}^p \frac{\Gamma\{\frac{1}{2}(n+1-j)+t+u\}}{\Gamma\{\frac{1}{2}(n+1-j)+u\}} \frac{\Gamma\{\frac{1}{2}(n-j)+u\}}{\Gamma\{\frac{1}{2}(n-j)\}}. \quad (41.91)$$

$$\begin{aligned} E\{|d|/|m|\}^u &= \prod_{j=1}^p \frac{\Gamma\{\frac{1}{2}(n+1-j)\} \Gamma\{\frac{1}{2}(n-j)+u\}}{\Gamma\{\frac{1}{2}(n+1-j)+u\} \Gamma\{\frac{1}{2}(n-j)\}} \\ &= \frac{\Gamma(\frac{1}{2}n) \Gamma\{\frac{1}{2}(n-p)+u\}}{\Gamma(\frac{1}{2}n+u) \Gamma\{\frac{1}{2}(n-p)\}} \\ &= \frac{B\{\frac{1}{2}(n-p)+u, \frac{1}{2}p\}}{B\{\frac{1}{2}(n-p), \frac{1}{2}p\}}. \end{aligned} \quad (41.92)$$

This is the u th moment of

$$dF = \frac{1}{B\{\frac{1}{2}(n-p), \frac{1}{2}p\}} x^{\frac{1}{2}(n-p-2)} (1-x)^{\frac{1}{2}(p-2)} dx, \quad (41.94)$$

which is uniquely determined by its moments. This then is the distribution of the ratio $|d|/|m|$, and on substitution for T from (41.82) we arrive at (41.86).

It will be seen that the essential feature of the T^2 statistic is that it is of form $\mathbf{z}'\hat{\mathbf{V}}^{-1}\mathbf{z}$, where \mathbf{z} is a multinormal vector with means zero and dispersion matrix \mathbf{V} , while $\hat{\mathbf{V}}$ is the ML estimator of \mathbf{V} , adjusted to be unbiased. Above, we had $\mathbf{z} = \bar{\mathbf{x}}$, $\hat{\mathbf{V}} = \mathbf{d}/\{n(n-1)\}$. Similarly, we may define a test statistic for the two-sample case (generalizing "Student's" two-sample t^2 test) with $\mathbf{z} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, $\hat{\mathbf{V}} = \mathbf{P}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$, where \mathbf{P} is the unbiased estimator of the common dispersion matrix of the two populations, calculated from the pooled samples.

41.18 So far we have generalized to p dimensions the normal distribution theory of means, dispersions, and the "Student" ratio. It is natural to enquire whether there is a similar generalization of the Fisher test for the ratio of two independent variances. We defer a discussion of this topic to the next chapter, where it arises naturally in connection with tests of hypotheses. We may remark here, however, that exact distributions in closed form corresponding to Fisher's z , for example, are difficult to derive, but that moments can usually be obtained in the manner of 41.12. A further point of some interest is that, in generalizations of variance analysis, it is not the ratio of two independent dispersion determinants which arises for test, but the

ratio of type $|c_1|/|c_1+c_2|$ where c_1 and c_2 are independent. In the univariate case we can transform simply from s_1^2/s_2^2 to $s_1^2/(s_1^2+s_2^2)$, but this is no longer possible in two or more dimensions.

Large-sample results

41.19 Even for normal variation, where we are not concerned with cumulants of order higher than the second, there is an embarrassing profusion of parameters, $\frac{1}{2}p(p+3)$ in all. For means, p variances, and $\frac{1}{2}p(p-1)$ covariances (or correlations), $\frac{1}{2}p(p+3)$ in all. For p as low as 5 there are 20, and for $p = 10$ there are 65. The distributional problems are correspondingly intractable unless we assume away many of these parameters, e.g. by considering the case of independent variables, for which all correlations vanish. In such circumstances large-sample results are not to be despised, though they are often neglected. To the first order in n we have

$$E(c_{jk}) = \gamma_{jk}. \quad (41.95)$$

To the same order

$$E(c_{jk}c_{lm}) = \frac{1}{n^2} E \left\{ \sum_{\alpha=1}^n x_{j\alpha} x_{k\alpha} \sum_{\beta=1}^n x_{l\beta} x_{m\beta} \right\}. \quad (41.96)$$

If $\alpha \neq \beta$ the two sums on the right are independent. There are $n(n-1)$ such cases and the expectation is $\gamma_{jk}\gamma_{lm}$. If $\alpha = \beta$ there are n terms such as $E(x_{j\alpha}x_{k\alpha}x_{l\alpha}x_{m\alpha})$. In general this involves fourth-order cumulants, but for normal variation we see from the c.f. of the x 's,

$$\phi = \exp \left(-\frac{1}{2} \sum_j \sum_k \gamma_{jk} t_j t_k \right),$$

that

$$E(x_{j\alpha}x_{k\alpha}x_{l\alpha}x_{m\alpha}) = \gamma_{jk}\gamma_{lm} + \gamma_{jm}\gamma_{kl} + \gamma_{jl}\gamma_{km}. \quad (41.97)$$

Substitution in (41.96) then gives us

$$E(c_{jk}c_{lm}) = \gamma_{jk}\gamma_{lm} + \frac{1}{n}(\gamma_{jm}\gamma_{kl} + \gamma_{jl}\gamma_{km})$$

and hence

$$\text{cov}(c_{jk}, c_{lm}) = \frac{1}{n}(\gamma_{jm}\gamma_{kl} + \gamma_{jl}\gamma_{km}). \quad (41.98)$$

In particular, if $j = k = l = m = 1$ we have the known result

$$\text{var } c_{11} = \frac{2\gamma_{11}^2}{n} = \frac{2\sigma_1^4}{n}. \quad (41.99)$$

Example 41.6

In our notation, it being indifferent whether we write parent or sample symbols,

$$\begin{aligned} r_{12} &= \frac{\gamma_{12}}{(\gamma_{11}\gamma_{22})^{\frac{1}{2}}}, \\ \frac{dr_{12}}{r_{12}} &= \frac{d\gamma_{12}}{\gamma_{12}} - \frac{1}{2} \frac{d\gamma_{11}}{\gamma_{11}} - \frac{1}{2} \frac{d\gamma_{22}}{\gamma_{22}}. \end{aligned} \quad (41.100)$$

Likewise

$$\frac{dr_{13}}{r_{13}} = \frac{d\gamma_{13}}{\gamma_{13}} - \frac{1}{2} \frac{d\gamma_{11}}{\gamma_{11}} - \frac{1}{2} \frac{d\gamma_{33}}{\gamma_{33}}. \quad (41.101)$$

Multiplying (41.100) and (41.101) and taking expectations, we have

$$\begin{aligned} \frac{\text{cov}(r_{12}, r_{13})}{r_{12} r_{13}} = & \frac{\text{cov}(\gamma_{12}, \gamma_{13})}{\gamma_{12} \gamma_{13}} - \frac{1}{2} \frac{\text{cov}(\gamma_{11}, \gamma_{12})}{\gamma_{11} \gamma_{12}} - \frac{1}{2} \frac{\text{cov}(\gamma_{11}, \gamma_{13})}{\gamma_{11} \gamma_{13}} \\ & - \frac{1}{2} \frac{\text{cov}(\gamma_{22}, \gamma_{13})}{\gamma_{22} \gamma_{13}} - \frac{1}{2} \frac{\text{cov}(\gamma_{33}, \gamma_{12})}{\gamma_{33} \gamma_{12}} + \frac{1}{4} \frac{\text{var } \gamma_{11}}{\gamma_{11}^2} \\ & + \frac{1}{4} \frac{\text{cov}(\gamma_{11}, \gamma_{33})}{\gamma_{11} \gamma_{33}} + \frac{1}{4} \frac{\text{cov}(\gamma_{11}, \gamma_{22})}{\gamma_{11} \gamma_{22}} + \frac{1}{4} \frac{\text{cov}(\gamma_{22}, \gamma_{33})}{\gamma_{22} \gamma_{33}}. \end{aligned} \quad (41.102)$$

Relations of type (41.98) reduce this to

$$n \text{ cov}(r_{12}, r_{13}) = r_{23}(1 - r_{12}^2 - r_{13}^2) + \frac{1}{2} r_{12} r_{13} (r_{12}^2 + r_{13}^2 + r_{23}^2 - 1). \quad (41.103)$$

The correlation of r_{12} and r_{13} is easily determined from the consideration that $n \text{ var } r = (1 - r^2)^2$.

The latent roots of a dispersion matrix

41.20 In later chapters we shall encounter several situations in which we are interested in the latent roots of a stochastic matrix. If (c_{jk}) is a dispersion matrix we shall wish to study the behaviour of the roots of the p -ic in λ

$$|c_{jk} - \lambda \delta_{jk}| \equiv |\mathbf{c} - \lambda \mathbf{I}| = 0, \quad (41.104)$$

where δ_{jk} is the Kronecker symbol, equal to zero unless $j = k$, in which case it is equal to unity.

We take from matrix theory the result that if \mathbf{c} is a positive definite matrix the latent roots are all real and non-negative. Only exceptionally will the roots be equal, and if q of them are zero \mathbf{c} is singular and has rank $p - q$.

41.21 In point of fact (41.104) is a particular case of a rather more general form

$$|c_{jk} - \lambda b_{jk}| = 0, \quad (41.105)$$

where \mathbf{b} , \mathbf{c} are independent dispersion matrices based on m , n observations respectively.

We may write equivalently

$$|u(c_{jk} + b_{jk}) - b_{jk}| = 0 \quad (41.106)$$

with

$$u = \frac{1}{1 + \lambda}, \quad \lambda = \frac{1 - u}{u}. \quad (41.107)$$

The complexity of the distributional problem arises from the fact that the roots in λ or u are not algebraic functions of the dispersions. It is easier to derive sampling distributions of symmetric functions of the roots than of the individual roots themselves.

41.22 We assume that the parent variation is normal with unit variances and zero covariances. The joint distribution of the c_{jk} and b_{jk} then has the frequency function, as at (41.58),

$$f \propto \frac{|c|^{\frac{1}{2}(n-p-2)} |b|^{\frac{1}{2}(m-p-2)}}{\prod_{j=1}^p \Gamma\{\frac{1}{2}(n-j)\} \Gamma\{\frac{1}{2}(m-j)\}} \exp \left\{ -\frac{1}{2} n \sum_{j=1}^p (c_{jj} + b_{jj}) \right\}. \quad (41.108)$$

In Chapter 43 we shall see that to each root u_j of (41.106) there corresponds a latent vector \mathbf{l}_j such that

$$\{\mathbf{b} - u_j(\mathbf{b} + \mathbf{c})\}\mathbf{l}_j = 0 \quad (41.109)$$

and we may choose the \mathbf{l}_j so that

$$\mathbf{l}_j'(\mathbf{b} + \mathbf{c})\mathbf{l}_j = 1. \quad (41.110)$$

If \mathbf{l} is the $p \times p$ matrix of latent vectors we have from (41.109)

$$\mathbf{b}\mathbf{l} = (\mathbf{b} + \mathbf{c})\mathbf{l}\mathbf{u} \quad (41.111)$$

where \mathbf{u} is a diagonal matrix whose elements are u_1, u_2, \dots, u_p . We now suppose that the u 's are arranged in descending order of magnitude, $u_1 \geq u_2 \geq \dots \geq u_p$.

We also have from (41.109)

$$\mathbf{l}_k' \mathbf{b} \mathbf{l}_j = u_j \mathbf{l}_k' (\mathbf{b} + \mathbf{c}) \mathbf{l}_j \quad (41.112)$$

and by transposition (\mathbf{b} and \mathbf{c} being symmetric)

$$\mathbf{l}_k' \mathbf{b} = \mathbf{l}_k' u_k (\mathbf{b} + \mathbf{c})$$

so that

$$\mathbf{l}_k' \mathbf{b} \mathbf{l}_j = u_j \mathbf{l}_k' (\mathbf{b} + \mathbf{c}) \mathbf{l}_j. \quad (41.113)$$

It follows from (41.112) and (41.113) that, if $u_j \neq u_k$,

$$\mathbf{l}_k' (\mathbf{b} + \mathbf{c}) \mathbf{l}_j = 0. \quad (41.114)$$

Multiplying (41.111) on the left by \mathbf{l}' and using (41.110) we have

$$\mathbf{l}' \mathbf{b} \mathbf{l} = \mathbf{u}. \quad (41.115)$$

From (41.110) and (41.114),

$$\mathbf{l}'(\mathbf{b} + \mathbf{c})(\mathbf{l} - \mathbf{I}) = 0 \quad (41.116)$$

and hence

$$\mathbf{b} + \mathbf{c} = (\mathbf{l}')^{-1} \mathbf{l}^{-1}. \quad (41.117)$$

Likewise from (41.115)

$$\mathbf{b} = (\mathbf{l}')^{-1} \mathbf{u} \mathbf{l}^{-1} \quad (41.118)$$

and hence

$$\mathbf{c} = (\mathbf{l}')^{-1} (\mathbf{I} - \mathbf{u}) \mathbf{l}^{-1}. \quad (41.119)$$

41.23 Looking back to (41.108), we see that with the transformation (41.118–19) the frequency function is given by

$$f \propto \frac{\left\{ \prod_{j=1}^p (1 - u_j) \right\}^{\frac{1}{2}(n-p-2)} \left\{ \prod_{j=1}^p u_j \right\}^{\frac{1}{2}(m-p-2)}}{\prod \Gamma\left\{\frac{1}{2}(n-j)\right\} \prod \Gamma\left\{\frac{1}{2}(m-j)\right\}} \times \text{a function of } \mathbf{l}. \quad (41.120)$$

We now consider the Jacobian of the transformation. It will appear that the distribution of u 's is independent of that of the \mathbf{l} 's and the contribution from the latter may be factorized off.

We observe from (41.118) and (41.119) that there are $\frac{1}{2}p(p+1)$ variables in each of \mathbf{b} and \mathbf{c} , $p(p+1)$ in all; and p^2 variables in \mathbf{l} and p u 's, again making $p(p+1)$.

The Jacobian of \mathbf{b} , \mathbf{c} is the same as that of \mathbf{b} , $\mathbf{b} + \mathbf{c}$. Writing \mathbf{g} for \mathbf{l}^{-1} in (41.118) and (41.119), we then have to consider the Jacobian of the transformation

$$\mathbf{b} = \mathbf{g}' \mathbf{u} \mathbf{g} \quad (41.121)$$

$$\mathbf{b} + \mathbf{c} = \mathbf{g}' \mathbf{g}. \quad (41.122)$$

We prove that $u_1 - u_2$ is a factor of the Jacobian. Although the argument is general, it will be clearer if we write it explicitly for $p = 2$.
The matrix \mathbf{b} then becomes

$$\begin{pmatrix} g_{11}^2 u_1 + g_{21}^2 u_2 & g_{11} g_{12} u_1 + g_{21} g_{22} u_2 \\ g_{12} g_{11} u_1 + g_{22} g_{21} u_2 & g_{12}^2 u_1 + g_{22}^2 u_2 \end{pmatrix}. \quad (41.123)$$

$\mathbf{b} + \mathbf{c}$ is the same with the u 's put equal to unity.
For the Jacobian we have

$$\frac{\partial \{b_{11}, b_{12}, b_{22}, (b+c)_{11}, (b+c)_{12}, (b+c)_{22}\}}{\partial (u_1, u_2, g_{11}, g_{12}, g_{21}, g_{22})} = \begin{vmatrix} g_{11}^2 & g_{21}^2 & 2g_{11}u_1 & 0 & 2g_{21}u_2 & 0 \\ g_{11}g_{12} & g_{21}g_{22} & g_{12}u_1 & g_{11}u_1 & g_{22}u_2 & g_{21}u_2 \\ g_{12}^2 & g_{22}^2 & 0 & 2g_{12}u_1 & 0 & 2g_{22}u_2 \\ 0 & 0 & 2g_{11} & 0 & 2g_{21} & 0 \\ 0 & 0 & g_{12} & g_{11} & g_{22} & g_{21} \\ 0 & 0 & 0 & 2g_{12} & 0 & 2g_{22} \end{vmatrix}$$

If $u_1 = u_2$ we can subtract multiples of the bottom three rows from the top three to obliterate all except the first two terms in these rows, and the determinant vanishes.

It follows that every $u_j - u_k, j > k$, is a factor of the Jacobian. The product of these factors is of degree $\frac{1}{2}p(p-1)$ in the u 's. There can be no other terms involving u because the Jacobian can be of no higher degree.

Thus, for the u 's alone we have from (41.120)

$$dF = k \frac{\{\prod (1-u_j)\}^{\frac{1}{2}(n-p-2)} (\prod u_j)^{\frac{1}{2}(m-p-2)}}{\prod \Gamma\{\frac{1}{2}(n-j)\} \Gamma\{\frac{1}{2}(m-j)\}} \prod_{j < k} (u_j - u_k) \prod du_j, \quad (41.124)$$

where k is some constant. To evaluate it by explicit integration would be very difficult. The following indirect route suffices:

k arises from terms in the original density and the Jacobian involving p and $m+n$, but not m separately. Write it then as $k(p, m+n)$. Note that $|b| = \prod u_j$. If we increase m by $2t$ in (41.124) and integrate we have the t th moment of $|b|$ except for a factor $k(p, m+n+2t)$. After the manner of 41.12 we can find this moment, and there results

$$\frac{k(p, m+n+2t)}{k(p, m+n)} = \prod_{j=1}^p \frac{\Gamma\{\frac{1}{2}(m+n-1-j)+t\}}{\Gamma\{\frac{1}{2}(m+n-1-j)\}}. \quad (41.125)$$

It follows that

$$k(p, m+n) = K(p) \prod_{j=1}^p \Gamma\{\frac{1}{2}(m+n-1-j)\}, \quad (41.126)$$

where $K(p)$ is a function of p only. To evaluate it, make the substitution in (41.124) of $u_j = 2v_j/n$ and let n tend to infinity. Our distribution becomes

$$\begin{aligned} dF &= (\prod v_i)^{\frac{1}{2}(m-p-2)} \exp(-\sum v_j) \prod (v_j - v_k) \prod dv_j \\ &\times \prod \frac{\Gamma\{\frac{1}{2}(m+n-1-j)\} K(p)}{\Gamma\{\frac{1}{2}(m-j)\} \Gamma\{\frac{1}{2}(n-j)\}}. \end{aligned} \quad (41.127)$$

This may be evaluated by step-by-step substitutions of the type

$$\begin{aligned} v_1 &= w_1 \\ v_j &= w_j + v_1, \quad j > 1, \end{aligned}$$

and choosing m at each stage so that the coefficient in Πu_j vanishes, as we may since the result is independent of m . We find

$$\frac{K(p)}{2^{\frac{1}{2}p(p-1)}} \Pi \frac{\Gamma(p+1-j)}{\Gamma\{\frac{1}{2}(p-j)\}} = 1. \quad (41.128)$$

Use of the duplication formula (41.56) for the Gamma function gives us

$$K(p) = \frac{\pi^{\frac{1}{2}p}}{\Pi \Gamma\{\frac{1}{2}(p+1-j)\}}.$$

Assembling all the factors, we find finally for the distribution

$$dF = \pi^{\frac{1}{2}p} \prod_{j=1}^p \frac{\Gamma\{\frac{1}{2}(m+n-1-j)\} u_j^{\frac{1}{2}(m-p-2)} (1-u_j)^{\frac{1}{2}(n-p-2)}}{\Gamma\{\frac{1}{2}(m-j)\} \Gamma\{\frac{1}{2}(n-j)\} \Gamma\{\frac{1}{2}(p+1-j)\}} \prod_{j < k} (u_j - u_k) \Pi du_j, \quad (41.129)$$

a remarkable form discovered in 1939 by Fisher, P. L. Hsu, S. N. Roy, Girshick, and Mood, independently—cf. Mood (1951).

The distribution of the λ 's is, of course, given by a simple substitution

$$u = 1/(1+\lambda).$$

41.24 In the case of (41.104), when the matrix \mathbf{b} reduces to the identity matrix, a slightly different result is obtained. We will quote the result for the distribution of the roots λ in this case. The reader may care to run through the foregoing proof and modify it where necessary to obtain this result:

$$dF = \frac{\pi^{\frac{1}{2}p}}{2^{\frac{1}{2}p(n-1)}} \prod_{j=1}^p \frac{\lambda_j^{\frac{1}{2}(n-p-2)} \exp\left\{-\frac{1}{2} \sum_{j=1}^p \lambda_j\right\}}{\Gamma\{\frac{1}{2}(n-j)\} \Gamma\{\frac{1}{2}(p+1-j)\}} \prod_{j < k} (\lambda_j - \lambda_k) \Pi d\lambda_j, \quad (41.130)$$

where now the λ_j are in descending order.

Example 41.7

These distributions are very intractable except in simple cases. Let us consider the case when $p = 2$. From (41.130) we have

$$dF = \frac{\pi^{\frac{1}{2}}}{2^{n-1}} \frac{(\lambda_1 \lambda_2)^{\frac{1}{2}(n-4)} \exp\left\{-\frac{1}{2}(\lambda_1 + \lambda_2)\right\}}{\Gamma\{\frac{1}{2}(n-1)\} \Gamma\{\frac{1}{2}(n-2)\}} (\lambda_1 - \lambda_2) d\lambda_1 d\lambda_2. \quad (41.131)$$

The duplication formula (41.56) for Gamma functions reduces the frequency function to

$$dF = \frac{(\lambda_1 \lambda_2)^{\frac{1}{2}(n-4)} \exp\left\{-\frac{1}{2}(\lambda_1 + \lambda_2)\right\} (\lambda_1 - \lambda_2) d\lambda_1 d\lambda_2}{4\Gamma(n-2)}. \quad (41.132)$$

If we try to integrate for λ_2 over the range 0 to λ_1 , we obtain for λ_1 an Incomplete Gamma function. On the other hand, for the functions

$$x = \lambda_1 \lambda_2, \quad y = \lambda_1 + \lambda_2$$

we find $1/J = |\lambda_1 - \lambda_2|$ and the distribution becomes

$$dF = \frac{x^{\frac{1}{2}(n-4)} e^{-\frac{1}{2}y} dx dy}{4\Gamma(n-2)}, \quad 0 \leq x \leq \frac{1}{4}y^2, \quad (41.133)$$

and on integration for x ,

$$dF = \frac{(\frac{1}{2}y)^{n-2} e^{-\frac{1}{2}y} d(\frac{1}{2}y)}{\Gamma(n-1)}, \quad 0 \leq y \leq \infty. \quad (41.134)$$

In fact, in this case the determinant reduces to

$$\begin{vmatrix} s_1^2 - \lambda & s_1 s_2 r \\ s_1 s_2 r & s_2^2 - \lambda \end{vmatrix} = \lambda^2 - (s_1^2 + s_2^2)\lambda + s_1^2 s_2^2 (1 - r^2) = 0. \quad (41.135)$$

The sum of the roots is thus equal to the sum of two independent variables and has a χ^2 distribution with $2n-2$ degrees of freedom.

41.25 We shall discuss the large-sample theory of latent roots in Chapter 43.

Non-central distributions

41.26 Just as for univariate χ^2 , t , and F (variance ratio), so there arise for study here non-central multivariate distributions, especially in the consideration of power-functions of tests based on T^2 or related statistics. As might be expected, the resulting distributions are very cumbersome. We may note particularly

- (a) The non-central Wishart distribution, as to which see T. W. Anderson (1946) and subsequent papers and his book (1958);
- (b) Non-central T^2 . Since T^2 is distributed in the F form this is effectively a non-central F —see 42.22 below.

41.27 In conclusion we may note some points which we shall have no space to develop in detail.

- (1) The distribution of latent roots (41.129) reduces for $p = 1$ to a Beta distribution. Foster and Rees (1957–8) therefore called it a “generalized Beta distribution.” Following a method due to S. N. Roy (1945) and Pillai (1956), they tabulated percentiles of the largest root for $p = 2, 3, 4$ and 5. Pillai (1966) has improved the method and tabulated (1964) up to $p = 7$.
- (2) Wagle (1962) approached the distribution problem by sampling experiments on an electronic computer. The task is not a light one, but results for $p = 2, 3, 4$, for all latent roots, were successfully obtained, and calculations for higher values are only a matter of machine time.
- (3) The Indian school, starting with some work by Mahalanobis (1930) on racial likeness, has developed some interesting work based on what is known as the D^2 -statistic. See, for example, R. C. Bose (1936), R. C. Bose and S. N. Roy (1938), and many later papers by S. N. Roy. The statistic may be defined as

$$D^2 = \sum_{j,k=1}^p a_{jk} (\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1k} - \bar{x}_{2k}) \quad (41.136)$$

where two samples, x_1 and x_2 , are drawn from two p -variate populations and (a_{jk}) is the inverse of the pooled dispersion matrix.

The corresponding parameter

$$\Delta^2 = \sum \alpha_{jk} (\mu_{1j} - \mu_{2j})(\mu_{1k} - \mu_{2k}) \quad (41.137)$$

is sometimes known as Mahalanobis' generalized distance.

In fact, D^2 is a simple function of Hotelling's T^2 defined for the two-sample case as in 41.17 above, i.e. $D^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) T^2$.

- (4) If $\mathbf{c}_1, \mathbf{c}_2$ follow a Wishart distribution based on p variables with sample numbers n_1 and n_2 , there exists a lower triangular matrix \mathbf{V} such that

$$\mathbf{c}_1 + \mathbf{c}_2 = \mathbf{V}\mathbf{V}'$$

(cf. Exercise 41.16). If \mathbf{L} is defined by

$$\mathbf{c}_1 = \mathbf{V}\mathbf{L}\mathbf{V}',$$

then \mathbf{V} and \mathbf{L} are independently distributed and \mathbf{L} has frequency function

$$f \propto |\mathbf{L}|^{\frac{1}{2}(n_1-p-2)} |\mathbf{I} - \mathbf{L}|^{\frac{1}{2}(n_2-p-2)}.$$

This result is originally due to Hsu (1939). See also Kshirsagar (1961). The distribution of \mathbf{L} is sometimes called "the multivariate Beta distribution."

- (5) A summary of work on latent roots is given by A. T. James (1964). See also A. T. James (1966) and Pillai (1966).

EXERCISES

41.1 If $\bar{\mathbf{x}}$ is the mean of a p -variate sample from a normal population with mean μ and dispersion matrix γ show that

$$n(\bar{\mathbf{x}} - \mu_0)' \gamma^{-1} (\bar{\mathbf{x}} - \mu_0)$$

is distributed as non-central χ^2 with p degrees of freedom and non-central parameter

$$n(\mu - \mu_0)' \gamma^{-1} (\mu - \mu_0)$$

where μ_0 is a given vector.

(R. C. Bose, 1936)

41.2 $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are the sample means from two p -variate normal populations with means μ_1, μ_2 and common dispersion matrix γ . If $\mathbf{y} = \mathbf{x}_1 - \mathbf{x}_2$ and $\mathbf{v} = \mu_1 - \mu_2$ show that

$$\frac{n_1 n_2}{n_1 + n_2} (\mathbf{y} - \mathbf{v})' \gamma^{-1} (\mathbf{y} - \mathbf{v}) \geq \chi^2$$

is a confidence region for \mathbf{v} , n_1 and n_2 being the respective sample numbers.

41.3 In 41.4, show that $\bar{\mathbf{x}}$ and the sample matrix \mathbf{a} ($= \mathbf{c}^{-1}$) are sufficient statistics for μ and γ^{-1} .

41.4 In a four-variate normal distribution show that the correlation between the covariances c_{12} and c_{34} is

$$\frac{\rho_{13} \rho_{24} + \rho_{14} \rho_{23}}{\{(1 + \rho_{12}^2)(1 + \rho_{34}^2)\}^{1/2}}$$

(Wishart, 1929)

41.5 A multivariate normal population has means μ , all variances equal to σ^2 and all correlations equal to ρ . Defining

$$(n-1)ps^2 = \sum_{i,j} (x_{ij} - \bar{x}_i)^2,$$

$$(n-1)(p-1)ps^2 r = \sum_{i \neq k} \sum_j (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k),$$

show that the joint frequency of

$$u = s^2 \{1 + (p-1)r\}$$

$$v = s^2 (1-r)$$

is given by

$$dF \propto u^{\frac{1}{2}(n-3)} v^{\frac{1}{2}(p-1)(n-1)-1} \exp \left\{ \frac{1}{2}(n-1) \left(\frac{u}{\alpha} + \frac{(p-1)v}{\theta} \right) \right\} du dv$$

where

$$\alpha = \sigma^2 \{1 + (p-1)\rho\},$$

$$\theta = \sigma^2 (1-\rho).$$

Hence show that $u\theta/(v\alpha)$ is distributed in the F form with $n-1$, $(p-1)(n-1)$ d.fr., and derive confidence intervals for ρ .

(Geisser, 1964)

41.6 \mathbf{c} is distributed in the Wishart form in samples from a population with dispersion matrix γ . Show that $\mathbf{h}'\mathbf{c}\mathbf{h}$ is distributed with corresponding parameter $\mathbf{h}'\gamma\mathbf{h}$, \mathbf{h} being an arbitrary non-singular $p \times p$ matrix.

41.7 If \mathbf{x} is distributed $N(\mathbf{0}, \gamma)$, i.e. normal with zero mean and dispersion matrix γ , and \mathbf{M} is an orthogonal matrix, show that $\mathbf{y} = \mathbf{M}\mathbf{x}$ is also distributed $N(\mathbf{0}, \gamma)$. In particular if \mathbf{M} is the Helmert matrix of 41.6 show that \mathbf{c} may be represented as

$$\sum_{l=1}^{n-1} \mathbf{y}_l' \mathbf{y}_l$$

where the \mathbf{y} 's are independent and distributed as $N(\mathbf{0}, \gamma)$. Deduce the additive property of Wishart matrices of 41.10.

41.8 With reference to Example 41.3 show that the frequency function of $\{|\mathbf{c}|/|\gamma|\}^{1/p}$, say y , is given approximately by

$$\frac{\alpha^{\frac{1}{2}p(n-p)} \gamma^{\frac{1}{2}p(n-p)-1} e^{-\alpha y}}{\Gamma \left\{ \frac{1}{2}p(n-p) \right\}},$$

where

$$\alpha = np \left\{ 1 - \frac{(p-1)(p-2)}{2n} \right\}^{1/p}.$$

(Hoel, 1937)

41.9 Show that for a sample dispersion matrix \mathbf{c} , $n^{\frac{1}{2}}\{|\mathbf{c}|/|\gamma| - 1\}$ is distributed about zero with variance $2p$ for large samples.

(T. W. Anderson, 1958)

41.10 If a sample of n is chosen from a p -variate normal population, the variates being grouped into k classes $x_1, \dots, x_{p_1}; x_{p_1+1}, \dots, x_{p_1+p_2}; \dots; x_{p_1+p_2+\dots+p_{k-1}+1}, \dots, x_p$, consider the function

$$W = \frac{|\mathbf{r}|}{|\mathbf{r}_{ij}^{(0)}|}$$

where $r_{ii} = 1$ and $r_{ij}^{(0)}$ is zero if the variates belong to different classes and equals the correlation r_{ij} if they belong to the same class.

Show that the LR test of the independence of the k sets of variates is

$$l = W^{\frac{1}{2}n}$$

and show that

$$\mu'_r(W) = \prod_{t=1}^k \prod_{j=1}^{p_t} \left\{ \frac{\Gamma\{\frac{1}{2}(n-j)\}}{\Gamma\{\frac{1}{2}(n-j)+r\}} \right\} \prod_{j=1}^p \frac{\Gamma\{\frac{1}{2}(n-j)+r\}}{\Gamma\{\frac{1}{2}(n-j)\}}.$$

(Wilks, 1935)

41.11 As a particular case of the last exercise, show that if a single variate x_1 is independent of a second set x_2, \dots, x_p , then

$$\mu'_r(W) = \frac{\Gamma\{\frac{1}{2}(n-1)\} \Gamma\{\frac{1}{2}(n-p)+r\}}{\Gamma\{\frac{1}{2}(n-1)+r\} \Gamma\{\frac{1}{2}(n-p)\}},$$

and hence find the distribution of the multiple correlation coefficient when the parent coefficient is zero.

(Wilks, 1935)

41.12 Show algebraically that Hotelling's T^2 is invariant under linear transformations of the p variates.

41.13 For a pair of normal variates with correlation ρ , show that, defining v by

$$v = \frac{nc_{12}}{\sigma_1 \sigma_2 (1-\rho^2)},$$

we have for the frequency function of v

$$f(v) = \frac{(1-\rho^2)^{\frac{1}{2}(n-1)} e^{pv}}{\pi^{\frac{1}{2}} 2^{\frac{1}{2}n-1} \Gamma\{\frac{1}{2}(n-1)\}} \{v^{\frac{1}{2}n-1} K_{\frac{1}{2}n-1}(v)\},$$

for $v > 0$ and a similar expression with $-v$ for v inside the curly brackets if $v < 0$. Here K is the Bessel function of second kind with imaginary argument.

(Wishart and Bartlett, 1933)

41.14 In equation (41.129) with $p = 2$ show that the distribution of $y = (1-u_1)(1-u_2)$ is given by

$$dF = \frac{(1-\sqrt{y})^{n-p} (\sqrt{y})^{m-p-1} d\sqrt{y}}{B(m-p, n-p+1)}.$$

41.15 Verify equation (41.62).

41.16 (Bartlett decomposition.) Let x_{jk} , $j = 1, 2, \dots, p$; $k = 1, 2, \dots, n$, be independent $N(0, 1)$ variables. Take

$$y_1 = x_1$$

$$y_2 = x_2 - b'_{21} y_1$$

$$\vdots$$

$$y_p = x_p - b'_{p1} y_1 - \dots - b'_{p,p-1} y_{p-1}$$

and take the y 's to be orthogonal so that $y'_j y_k = 0$, $j \neq k$. Then $b'_{jk} = y'_k x_j / y'_k y_k$. Take $b_{jk} = (y'_k y_k)^{\frac{1}{2}} b'_{jk}$. Show that

where B is the triangular matrix

$$x'x = BB'$$

$$\begin{pmatrix} (y'_1 y_1)^{\frac{1}{2}} & 0 & 0 & \dots & 0 \\ b_{21} & (y'_2 y_2)^{\frac{1}{2}} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{p1} & b_{p2} & \vdots & \dots & (y'_p y_p)^{\frac{1}{2}} \end{pmatrix}.$$

Show that the b_{jk} , $k = 1, \dots, (p-1)$ are independent $N(0, 1)$ variables and hence that each y_k/y_j is a χ^2 variable with $n-k+1$ degrees of freedom, independent of the b 's and the other y_j/y_j .

(Bartlett, 1933. Cf. Kshirsagar, 1959)

41.17 In the foregoing exercise, by taking determinants in the equation $\mathbf{x}'\mathbf{x} = \mathbf{B}\mathbf{B}'$, show that y_k/y_j is the ratio of two product-sum determinants. Hence show that the diagonal elements of the inverse of $\mathbf{x}'\mathbf{x}$ are distributed as the reciprocal of a χ^2 variable with $n-p+1$ degrees of freedom.

(Wijsman, 1957; see also Kshirsagar, 1959)

41.18 Use the previous exercise to prove the result of 41.13, that $|d|/|\gamma|$ is the product of p independent χ^2 factors with degrees of freedom $n-1, n-2, \dots, n-p$.

(Wijsman, 1957)

41.19 Verify the result (41.130).

41.20 From (41.53) show that $|c|$ is a biased estimator of $|\gamma|$. Show also that the bias is not removed by dividing product-sums by $n-1$ instead of n to obtain covariances.

CHAPTER 42

TESTS OF HYPOTHESES IN MULTIVARIATE ANALYSIS

42.1 The exact theory of multivariate analysis, in the present state of knowledge, is concerned almost entirely with normal variation, and we have seen in the previous chapter that ML estimators of means and dispersions are the corresponding sample statistics. A general theory of estimation, other than the ordinary maximum-likelihood method, has yet to be produced for practical use. Methods based on Bayesian prior probabilities or fiducial arguments have been discussed in the literature but lead to some rather anomalous results on occasion (see, for example, Dempster, 1966). We shall content ourselves with the normal ML estimators, which in any case have a certain plausibility. The only point to comment upon concerns bias.

42.2 Consider, for example, the dispersion determinant $|c|$ as an estimator of the parent $|\gamma|$. From equation (41.53) with $t = 1$ we find

$$E|c| = \prod_{j=1}^p \left(1 - \frac{j}{n}\right) |\gamma|. \quad (42.1)$$

To order n^{-1} the multiplicative factor on the right is

$$1 - \frac{1}{n} \sum_{j=1}^p j = 1 - \frac{p(p+1)}{2n}. \quad (42.2)$$

The bias may therefore be appreciable. We may, however, remove it, or at least reduce it to order n^{-2} , either by multiplying $|c|$ by the reciprocal of the right-hand side of (42.2) or by the device due to Quenouille (17.10, Vol. 2). If $|c|_n$ represents the dispersion determinant based on n observations and $|c|_{n-1}$ the similar determinant based on $n-1$, we construct the estimator

$$\text{Est } |\gamma| = n |c|_n - (n-1) \text{Av } |c|_{n-1} \quad (42.3)$$

where the average on the right is taken over all the n possible determinants obtained by dropping one observation. We then have, to order n^{-1} ,

$$\begin{aligned} \frac{E\{\text{Est } |\gamma|\}}{|\gamma|} &= n \left\{1 - \frac{p(p+1)}{2n}\right\} - (n-1) \left\{1 - \frac{p(p+1)}{2(n-1)}\right\} \\ &= 1, \end{aligned}$$

and the estimator is unbiased to order n^{-1} . The idea is quite straightforward; the difficulty in applying it lies in the amount of calculation involved in computing all the dispersion determinants, though this is not insuperable for an electronic computer.

Homogeneity tests

42.3 A natural generalization of the variance analysis considered in Chapter 35 arises if we consider samples from k different p -variate populations and enquire whether

the parents may be identical. There are, as usual, three types of hypothesis to consider:

- H : that the populations have the same means and dispersions, namely are identical;
- H_1 : that the populations have the same dispersions but may differ in the means;
- H_2 : it is known that the populations have the same dispersions; the hypothesis is that they have the same means.

There are, of course, hybrid hypotheses, e.g. given certain dispersions but not others, that the others are equal.

42.4 For testing simple hypotheses, the Neyman-Pearson lemma of 22.10 (Vol. 2) applies to multivariate distributions without change. Similarly the likelihood-ratio method of Chapter 24, with the same plausibility, may be adopted as a test statistic for composite hypotheses.

One property of maximization procedures is worth noticing. If we are maximizing, say $f(\theta_1, \theta_2)$ for variations in θ_1 and θ_2 , we may solve the simultaneous equations

$$\frac{\partial f}{\partial \theta_1} = 0, \quad \frac{\partial f}{\partial \theta_2} = 0. \quad (42.4)$$

It is, however, equivalent to solve $\partial f / \partial \theta_1 = 0$ for θ_1 , substitute in f , and then solve $df / d\theta_2 = 0$.

42.5 Consider, then, k multivariate normal populations with means typified by μ_{jt} ($j = 1, 2, \dots, p$; $t = 1, 2, \dots, k$) and dispersions by γ_{jlt} or equivalently $\sigma_{jt} \sigma_{lt} \rho_{jlt}$. Let there be a sample of n_t from the t th population. If α_{jlt} is inverse to γ_{jlt} the likelihood function of all samples together is

$$\prod_{t=1}^k \frac{|\alpha_t|^{\frac{1}{2}n_t}}{(2\pi)^{\frac{1}{2}pn_t}} \exp \left\{ -\frac{1}{2} \sum_{t=1}^k S \sum_{j,l=1}^p \alpha_{jlt} (x_{jt} - \mu_{jt})(x_{lt} - \mu_{lt}) \right\}. \quad (42.5)$$

If all μ 's and γ 's are equal, the corresponding likelihood is

$$\frac{|\alpha|^{\frac{1}{2}n}}{(2\pi)^{\frac{1}{2}pn}} \exp \left\{ -\frac{1}{2} S \sum_{j,l=1}^p \alpha_{jl} (x_j - \mu_j)(x_l - \mu_l) \right\}, \quad (42.6)$$

where

$$n = \sum_{t=1}^k n_t. \quad (42.7)$$

In accordance with the usual procedure (Chapter 24), we estimate the parameters in (42.5) by ML and substitute in it to obtain the unconditioned maximum L_1 . Likewise for (42.6) to obtain the conditioned maximum L_0 . We then use the ratio $l = L_0/L_1$, or some monotonic function of it, as the test criterion.

The logarithm of the likelihood (42.5) becomes the sum of k terms which, being independent, can be maximized separately. We find, as expected,

$$\hat{\mu}_{jt} = \bar{x}_{jt} \quad (42.8)$$

$$\hat{\alpha}_{jlt} = a_{jlt} \quad (42.9)$$

$$\hat{\gamma}_{jlt} = c_{jlt}. \quad (42.10)$$

Substitution in the exponential term in (42.5) yields a constant, for

$$\sum_{j,l} a_{jll} c_{jll} = 1. \quad (42.11)$$

Thus, except for a constant,

$$L_1 = \prod_{t=1}^k \frac{1}{|c_{jll}|^{\frac{1}{2}n_t}}. \quad (42.12)$$

Likewise from (42.6) we obtain

$$L_0 = \frac{1}{|c_{jll}|^{\frac{1}{2}n}}, \quad (42.13)$$

where c_{jll} is the dispersion for all k samples pooled together. The test criterion is then given by

$$\begin{aligned} l_H = \frac{L_0}{L_1} &= \frac{\prod |c_{jll}|^{\frac{1}{2}n_t}}{\prod |c_{jll}|^{\frac{1}{2}n}} \\ &= \prod_{t=1}^k \left\{ \frac{|c_{jll}|}{|c_{jll}|} \right\}^{\frac{1}{2}n_t}. \end{aligned} \quad (42.14)$$

As in the univariate case, l_H may vary from 0 to 1. The nearer to unity, the more we are inclined to accept the hypothesis that all means and all dispersions are equal.

42.6 The same technique gives us tests for H_1 and H_2 . We quote the results without proof.

Let c_{jla} be the average dispersion taken over the k populations, namely

$$c_{jla} = \frac{1}{n} \sum_{t=1}^k \sum_{u=1}^{n_t} (x_{ju} - \bar{x}_j)(x_{lu} - \bar{x}_l). \quad (42.15)$$

Then for H_1 ,

$$l_{H_1} = \prod_{t=1}^k \left\{ \frac{|c_{jll}|}{|c_{jla}|} \right\}^{\frac{1}{2}n_t}. \quad (42.16)$$

For H_2 ,

$$l_{H_2} = \left\{ \frac{|c_{jla}|}{|c_{jll}|} \right\}^{\frac{1}{2}n}. \quad (42.17)$$

We note that, as in the univariate case (cf. Exercise 24.6),

$$l_H = l_{H_1} l_{H_2}. \quad (42.18)$$

Our test criteria thus appear as the ratios of dispersion determinants.

42.7 To apply the tests we require the distributions of the criteria. In a few cases they can be obtained explicitly. In all cases we can obtain moments after the manner of 41.12. For practical purposes, however, it is enough to rely on an approximation due to Wilks, to the effect that $-2 \log l$ is distributed as χ^2 with d.fr. equal to the number of constraints imposed by the hypothesis. The proof is a simple extension of that in 24.7 (Vol. 2).

We have left the criteria l in the form in which they naturally arise. Clearly any power of l would serve our purpose equally well. In particular, we might use the $(2/n)$ th power, in which case the criterion for H_2 in (42.17) becomes the ratio of determinants and it is $-n$ times the logarithm of this ratio which is distributed as χ^2 .

42.8 Consider now the moments of the LR criterion (42.14) for testing H . We have

$$\begin{aligned} c_{jl} &= \frac{1}{n} S (x_{jt} - \bar{x}_j)(x_{lt} - \bar{x}_l) \\ &= \frac{1}{n} \sum_{t=1}^k S (x_{jt} - \bar{x}_{jt})(x_{lt} - \bar{x}_{lt}) + \frac{1}{n} \sum_{t=1}^k n_t (\bar{x}_{jt} - \bar{x}_j)(\bar{x}_{lt} - \bar{x}_l) \\ &= c_{jla} + c_{jlm}, \end{aligned} \quad (42.19)$$

where c_{jlm} is the dispersion of sample means about pooled means. Following the device used in 41.17 we can write the likelihood of dispersions in two ways, one involving $|c|$ and the other involving $|c_{jla}|$ and $|c_{jlm}|$. We then find

$$\begin{aligned} E(l_H^r) &= \prod_{t=1}^k \left[\left(\frac{n}{n_t} \right)^{\frac{1}{2}prn_t} \prod_{j=1}^p \frac{\Gamma[\frac{1}{2}\{n_t(1+r)-j\}]}{\Gamma[\frac{1}{2}(n_t-j)]} \right. \\ &\quad \left. \times \prod_{j=1}^p \frac{\Gamma[\frac{1}{2}(n-j)]}{\Gamma[\frac{1}{2}\{n(1+r)-j\}]} \right]. \end{aligned} \quad (42.20)$$

This and the following results are due to Wilks (1932), to whom reference may be made for details.

In a similar way it may be shown that

$$\begin{aligned} E(l_{H_1}^r) &= \prod_{t=1}^k \left[\left(\frac{n}{n_t} \right)^{\frac{1}{2}prn_t} \prod_{j=1}^p \frac{\Gamma[\frac{1}{2}\{n_t(1+r)-j\}]}{\Gamma[\frac{1}{2}(n_t-j)]} \right. \\ &\quad \left. \times \prod_{j=1}^p \frac{\Gamma[\frac{1}{2}(n-k+1-j)]}{\Gamma[\frac{1}{2}\{n(1+r)-k+1-j\}]} \right]. \end{aligned} \quad (42.21)$$

$$E(l_{H_2}^r) = \prod_{j=1}^p \frac{\Gamma[\frac{1}{2}\{n(1+r)-k+1-j\}]}{\Gamma[\frac{1}{2}(n-k+1-j)]} \frac{\Gamma[\frac{1}{2}(n-j)]}{\Gamma[\frac{1}{2}\{n(1+r)-j\}]} \quad (42.22)$$

Note that as in Exercise 24.6 the moments of l_H are the product of the moments of the other two l 's. This implies that l_{H_1} and l_{H_2} are independent when H holds, which is what we might expect from the independence of means and dispersions.

42.9 In passing we may remark on a possible source of confusion. In our notation n is the sample number, not the degrees of freedom. The form of the frequency distribution as written, for example, at (42.5) contains the n 's only in the preliminary constants. If the exponent were reducible to a quadratic form which transformed to a sum of $p-q$ squares the appropriate preliminary constants would have $p-q$ instead of p . And if the sum over sample values were equivalent to $n-q$ instead of n values we should have $n-q$ instead of n in the constants. Whether this affects the exponents in (42.5) and (42.6) depends on how we define the dispersions. In our usage the divisor is always n . Some writers use $v_t = n_t - 1$ instead of our n_t and $v = n - k$ instead of our n , in defining dispersions. The reason for so doing is the one noticed in Example 24.6, Vol. 2—the test is nearer to being unbiased.

42.10 For $p = 1$ the distributions reduce, of course, to those already familiar in univariate theory (cf. Exercises 24.4–6). The reader may care to verify as an exercise that this is so.

For $p = 2$ we find from (42.22) for the moments of l_H ,

$$\mu'_r = \frac{\Gamma[\frac{1}{2}\{n(1+r)-k\}] \Gamma[\frac{1}{2}(n-1)]}{\Gamma[\frac{1}{2}(n-k)] \Gamma[\frac{1}{2}n(1+r)-1]} \frac{\Gamma[\frac{1}{2}\{n(1+r)-k-1\}] \Gamma[\frac{1}{2}(n-2)]}{\Gamma[\frac{1}{2}(n-k-1)] \Gamma[\frac{1}{2}n(1+r)-2]}. \quad (42.23)$$

Use of the duplication formula (41.56) for the Gamma function reduces this to

$$\begin{aligned} \mu'_r &= \frac{\Gamma\{n(1+r)-k-1\} \Gamma(n-2)}{\Gamma(n-k-1) \Gamma\{n(1+r)-2\}} \\ &= B\{n(1+r)-k-1, k-1\} / B\{k-1, n-k-1\}. \end{aligned} \quad (42.24)$$

The moments of $(l_H)^{1/n}$, namely $\{|c_{jla}|/|c_{jl}|\}^{1/n}$, are then those of

$$dF = \frac{1}{B\{k-1, n-k-1\}} (1-x)^{k-2} x^{n-k-2} dx. \quad (42.25)$$

If p is even, we can use the duplication formula for the Gamma function to reduce the moments of the l -criteria to products of Beta functions, and the criteria are revealed as the product of certain independent Type I variables. But this fact is not very useful in giving explicit closed form to the distribution functions.

42.11 The most useful results for testing hypotheses in practice are asymptotic expressions. Following the treatment by Box (1949), we shall develop a general method along the lines of Example 41.4 in the previous chapter. The method, in fact, is applicable to a wide range of criteria depending on likelihood ratios.

Consider a variable W with moments

$$E(W^t) = \text{constant} \cdot \frac{\left[\prod_{j=1}^k y_j^{y_j} \right]^t}{\left[\prod_{j=1}^m x_j^{x_j} \right]^t} \frac{\prod_{j=1}^m \Gamma\{x_j(1+t) + \xi_j\}}{\prod_{j=1}^k \Gamma\{y_j(1+t) + \eta_j\}} \quad (42.26)$$

where

$$\sum_{j=1}^m x_j = \sum_{j=1}^k y_j. \quad (42.27)$$

In our treatment x_j and y_j will be large, of the order of n , the total sample number, and we may write $O(n)$ indifferently for $O(x)$ or $O(y)$.

Now take

$$M = -2 \log W \quad (42.28)$$

and let us find the characteristic function of ρM , where ρ lies between 0 and 1 and is a scaling constant which we may later choose at convenience. Taking t as the dummy variable in the c.f., we have for ρM

$$\begin{aligned} \phi(t) &= E(\exp it\rho M) = E(W^{-2\rho it}) \\ &= \text{constant} \cdot \frac{\left[\prod_{j=1}^k (y_j^{y_j}) \right]^{-2\rho it}}{\left[\prod_{j=1}^m (x_j^{x_j}) \right]^{-2\rho it}} \frac{\prod_{j=1}^m \Gamma\{x_j(1-2\rho it) + \xi_j\}}{\prod_{j=1}^k \Gamma\{y_j(1-2\rho it) + \eta_j\}}. \end{aligned} \quad (42.29)$$

Putting now

$$(1-\rho)x_j = \beta_j, \quad (1-\rho)y_j = \varepsilon_j, \quad (42.30)$$

we have for the cumulant-generating function of ρM

$$\psi(t) = g(t) - g(0)$$

where

$$g(t) = 2\rho it \left[\sum_{j=1}^m x_j \log x_j - \sum_{j=1}^k y_j \log y_j \right] + \sum_{j=1}^m \log \Gamma \{ \rho x_j (1 - 2it) + \beta_j + \xi_j \} - \sum_{j=1}^k \log \Gamma \{ \rho y_j (1 - 2it) + \varepsilon_j + \eta_j \}. \quad (42.31)$$

We now use the expansion of the Gamma function (valid for complex values)

$$\log \Gamma(x+h) = \log \sqrt{(2\pi)} + (x+h-\frac{1}{2}) \log x - x - \sum_{j=1}^n (-1)^j \frac{B_{j+1}(h)}{j(j+1)x^j} + R_{n+1}(x) \quad (42.32)$$

where the B 's are Bernoulli polynomials of order unity (3.25, Vol. 1). We then find, on expanding (42.31) and the corresponding $g(0)$,

$$\psi(t) = -\frac{1}{2}f \log(1-2it) + \sum_{j=1}^{\infty} \omega_j \{(1-2it)^{-j} - 1\} \quad (42.33)$$

where

$$f = -2 \left\{ \sum_{j=1}^m \xi_j - \sum_{j=1}^k \eta_j - \frac{1}{2}(m-k) \right\}, \quad (42.34)$$

$$\omega_j = \frac{(-1)^{j+1}}{j(j+1)} \left\{ \sum_{i=1}^m \frac{B_{j+1}(\beta_i + \xi_i)}{(\rho x_i)^j} - \sum_{i=1}^k \frac{B_{j+1}(\varepsilon_i + \eta_i)}{(\rho y_i)^j} \right\}. \quad (42.35)$$

We must remember that, from (42.30), β and ε are of order n unless $1-\rho$ is small. For $\rho = 1$ we have $\omega_1 = O(n^{-1})$. Thus we have

$$\psi(t) = -\frac{1}{2}f \log(1-2it) + O(n^{-1}), \quad (42.36)$$

and hence, to this order, $-2 \log W$ is distributed as χ^2 with f degrees of freedom.

Taking the approximation one stage further, we find, since

$$B_2(x) = x^2 - x + \frac{1}{6},$$

$$\omega_1 = \frac{1}{2\rho} \left[\sum_{i=1}^m \left\{ \frac{(\beta_i + \xi_i)^2 - (\beta_i + \xi_i) + \frac{1}{6}}{x_i} \right\} - \sum_{i=1}^k \left\{ \frac{(\varepsilon_i + \eta_i)^2 - (\varepsilon_i + \eta_i) + \frac{1}{6}}{y_i} \right\} \right]$$

which, by use of (42.30), reduces to

$$\omega_1 = \frac{1}{2\rho} \left[-(1-\rho)f + \sum_{i=1}^m \frac{\xi_i^2 - \xi_i + \frac{1}{6}}{x_i} - \sum_{i=1}^k \frac{\eta_i^2 - \eta_i + \frac{1}{6}}{y_i} \right]. \quad (42.37)$$

If we now take ρ such that $\omega_1 = 0$ we have

$$\psi(t) = -\frac{1}{2}f \log(1-2it) + O(n^{-2}). \quad (42.38)$$

In general then there exists a constant ρ such that ρW is distributed as χ^2 with f degrees of freedom to order n^{-2} .

Box (1949) has pushed the investigation a good deal further, but, as we have remarked, the cruder approximation (42.36) is usually good enough for practical purposes. See also Lawley (1956b), whose work was summarized in 24.9, Vol. 2.

Example 42.1

Let us find the χ^2 approximations to the distribution of l_n , the moments of which

are given by (42.20). Comparison with (42.26) shows that they are of the required form with

$$k = p, \quad y_j = \frac{1}{2}n, \quad \eta_j = -\frac{1}{2}j; \quad j = 1, 2, \dots, p.$$

$$m = pk, \quad x_j = \frac{1}{2}n, \quad t = 1, 2, \dots, k; \quad \xi_j = -\frac{1}{2}j.$$

Our first approximation is that $-2 \log l$ is distributed as χ^2 with degrees of freedom given by (42.34), namely

$$f = -2\left\{\left(-\frac{1}{2}\right)k\frac{1}{2}p(p+1) - \left(-\frac{1}{2}\right)\frac{1}{2}p(p+1) - \frac{1}{2}(pk-p)\right\} \\ = \frac{1}{2}(k-1)p(p+3). \quad (42.39)$$

This is, in fact, the number of parameters in the likelihood (42.5) less the number in (42.6) i.e. the number of constraints imposed by the hypothesis.

For a second approximation we find from (42.37), $\omega_1 = 0$,

$$0 = -(1-\rho)\frac{1}{2}(k-1)p(p+3) + p\left(\sum \frac{1}{n_j} - \frac{1}{n}\right)\frac{2p^2+9n+12}{12}, \\ \rho = 1 - \left(\sum_{j=1}^k \frac{1}{n_j} - \frac{1}{n}\right)\frac{2p^2+9p+12}{6(k-1)(p+3)}. \quad (42.40)$$

In expressions of this kind, it should be remembered that, in our convention, n_j and n are sample numbers. As we noted above, results are sometimes quoted in the literature for criteria based on the degrees of freedom $\nu_j = n_j - 1$ and $\nu = n - k$. This does not affect (42.39) but makes a difference to the second term in (42.40). In this case ξ_j is $\frac{1}{2}(1-j)$ and $\eta_j = \frac{1}{2}(k-j)$ and the corresponding expression to (42.40) is

$$\rho = 1 - \left(\sum_{j=1}^k \frac{1}{\nu_j} - \frac{1}{\nu}\right)\frac{2p^2+3p-1}{6(k-1)(p+3)} + \frac{1}{\nu}\left(\frac{p-k+2}{p+3}\right). \quad (42.41)$$

Tests of independence

42.12 The set of tests which we proceed to develop are, with few exceptions, all based on the foregoing ideas: the deduction of a likelihood criterion, the ascertainment of its moments, and the approximation to a χ^2 test or something a little more refined. We need not spend too much time on the derivation of the details, which may be left for verification to the student.

First of all we consider a test of independence. Given, as usual, a sample of n from a p -variate normal population, and given a division of the variables into q subsets containing p_1, p_2, \dots, p_q variables, it is required to test the hypothesis that each subset is independent of the others. We shall be particularly interested in the cases $q = 2$ and $q = p$.

If the parent dispersion matrix γ is partitioned into q^2 components

$$\begin{pmatrix} \gamma_{11} & \gamma_{12} & \dots & \gamma_{1q} \\ \gamma_{21} & \gamma_{22} & \dots & \gamma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{q1} & \gamma_{q2} & \dots & \gamma_{qq} \end{pmatrix} \quad (42.42)$$

the hypothesis under test is that

$$\gamma_{jk} = 0, \quad j \neq k. \quad (42.43)$$

We find for the likelihood ratio criterion, the alternative against (42.43) being (42.42),

$$l_H = \frac{|c|^{\frac{1}{2}n}}{\prod_{j=1}^q |c_{jj}|^{\frac{1}{2}n}}, \quad (42.44)$$

where as usual (c_{jk}) is the sample dispersion matrix. We can write equivalently

$$l_H^{2/n} = \frac{|c|}{\prod_{j=1}^q |c_{jj}|}. \quad (42.45)$$

Under the hypothesis, l is independent of its denominator and $|c_{jj}|$ is independent of $|c_{kk}|$. We then find from (41.53)

$$E(l^r) = \frac{\prod_{j=1}^p \Gamma\{\frac{1}{2}(n-j) + \frac{1}{2}rn\} \prod_{k=1}^q \prod_{j=1}^{p_k} \Gamma\{\frac{1}{2}(n-j)\}}{\prod_{j=1}^p \Gamma\{\frac{1}{2}(n-j)\} \prod_{k=1}^q \prod_{j=1}^{p_k} \Gamma\{\frac{1}{2}(n-j) + \frac{1}{2}rn\}}. \quad (42.46)$$

$-2 \log l$ is distributed approximately as χ^2 with

$$f = \frac{1}{2}\{p(p+1) - \sum_{j=1}^q p_j(p_j+1)\} \quad (42.47)$$

d.fr. For the more accurate approximation we have

$$\rho = 1 - \frac{2(p^3 - \sum p_j^3) + 9(p^2 - \sum p_j^2)}{6n(p^2 - \sum p_j^2)}. \quad (42.48)$$

In the case $p_j = 1$, all j , when we are testing the independence of all p variables,

$$f = \frac{1}{2}p(p-1) \quad (42.49)$$

$$\rho = 1 - \{(2p+11)/6n\}. \quad (42.50)$$

The criterion in this case is the $\frac{1}{2}$ nth power of $|c|$ divided by the product of the diagonal elements, namely the variances; or, equivalently, the $\frac{1}{2}$ nth power of the correlation determinant.

T. W. Anderson (1958) gives some further details and references to particular cases worked out explicitly by Wilks (1935).

Daly (1940) and Narain (1950) showed that tests of independence based on determinantal ratios are completely unbiased.

Sphericity test

42.13 We next consider a test whether an observed dispersion matrix \mathbf{c} can have arisen from a population with a matrix proportional to a given matrix γ . Since γ is known, we can transform it by a linear transformation to the identity matrix. Writing \mathbf{c} for the corresponding transform of the observed matrix, we then have to test whether \mathbf{c} is proportional to $\sigma^2 \mathbf{I}$ where σ^2 is unknown. This is described, for obvious reasons, as a sphericity test (Mauchly, 1940). We now find for the criterion

$$l = \left[\frac{|c|}{\left\{ \frac{1}{p} \text{trace } \mathbf{c} \right\}^p} \right]^{\frac{1}{2}n}. \quad (42.51)$$

The moments of $l^{2/n}$ are given by

$$E(l^{2r/n}) = p^{rp} \frac{\Gamma\{\frac{1}{2}p(n-1)\}}{\Gamma\{\frac{1}{2}p(n-1) + pr\}} \prod_{j=1}^p \frac{\Gamma\{\frac{1}{2}(n-j) + r\}}{\Gamma\{\frac{1}{2}(n-j)\}} \quad (42.52)$$

and as usual $-n \log l^{2/n}$ is distributed as χ^2 with

$$f = \frac{1}{2}p(p+1) - 1. \quad (42.53)$$

For the second approximation

$$\rho = 1 - \frac{2p^2 + p + 2}{6p(n-1)}. \quad (42.54)$$

Gleser (1966) shows that the sphericity test is unbiased.

42.14 The homogeneity tests can be used to generalize to p dimensions the tests of variance-analysis of univariate theory. Whereas in the latter we are concerned to compare independent estimators of variance, in the former we compare generalized variances, that is to say, dispersion determinants.

Example 42.2

We consider a two-dimensional case ($p = 2$), following Pearson and Wilks (1933). Five samples are available, each of twelve members, of aluminium die-castings ($k = 5$, $n_t = 12$ for all t , $n = 60$). On each of the 60 specimens two measurements are taken, tensile strength (in 1000 lb per sq. inch) which we call x , and hardness (Rockwell E) which we call y . The data may be summarized as follows:

Sample number t	mean x s.d.		mean y s.d.		Correlation
	mean	s.d.	mean	s.d.	
1	33.399	2.565	68.49	10.19	0.683
2	28.216	4.318	68.02	14.49	0.876
3	30.313	2.188	66.57	10.17	0.714
4	33.150	3.964	76.12	11.18	0.715
5	34.269	2.715	69.92	9.88	0.805

We test first of all the hypothesis H_1 that dispersions are homogeneous. We have the following results:

t	Sums of squares about means		Sum of products about means	Generalized variances (dispersion determinants)	\log_{10} of generalized variances
	x	y			
1	78.948	1247.18			
2	223.695	2519.31	214.18	365.204	2.56254
3	57.448	1241.78	657.62	910.401	2.95923
4	187.618	1473.44	190.63	243.029	2.38566
5	88.456	1171.73	375.91	938.451	2.92741
			259.18	253.281	2.40360
TOTALS	636.165	7653.44	1697.52		13.28844

For the pooled variances and covariances about respective means we have

$$c_{11a} = 636.165/60 = 10.6028$$

$$c_{22a} = 127.5573$$

$$c_{12a} = 28.2920$$

$$|c_{jka}| = 552.018.$$

The criterion is then, from (42.16) in the form $l^{2/n}$, given by

$$\frac{2}{n} \log l = \frac{1}{5} \sum_{t=1}^5 \{\log |c_{jtu}| - |c_{jlu}|\}$$

$$= 1.914,73,$$

with logs taken to base 10, giving $l^{2/n} = 0.8217$. For a test we find $-n \log_e l^{2/n} = 11.78$. The number of degrees of freedom is $3(k-1)$, namely 12. The observed value is consistent with homogeneity, and we can now proceed to hypothesis H_2 , that the means are equal given equality of dispersions. For this, we require to apply (42.17) and hence to find the pooled variance about pooled means. The data are as follows:

Source	d.fr.	SS (x)	SS (y)	SP (xy)
Between samples	$k-1 = 4$	306.089	662.77	214.86
Within samples	$n-k = 55$	636.165	7653.44	1697.52
TOTAL	$n-1 = 59$	942.254	8316.21	1912.38

The pooled dispersion determinant is then 1160.77 and the criterion is given by

$$-60 \log (552.018/1160.77) = 44.59.$$

The number of degrees of freedom is $2(k-1) = 8$. The result rejects H_2 at extremely small test sizes.

We conclude that there is heterogeneity in the means. We now test x and y separately.

	Estimates of variance		d.fr.
	x	y	
Between samples	76.522	165.69	4
Within samples	11.566	139.15	55

An ordinary F -test shows that at the 1 per cent point the differences between tensile strength, but not the differences between hardness, are responsible for the heterogeneity.

Multivariate regression

42.15 We now suppose that our variables x are linearly related to a set of z 's which may be regarded as fixed, by

$$\underset{p \times n}{x} = \underset{p \times q}{\beta} \underset{q \times n}{z} + \underset{p \times n}{\epsilon}. \quad (42.55)$$

β is a $p \times q$ matrix of coefficients, and ϵ is a $p \times n$ matrix of errors. If its sub-vectors $\epsilon_1, \epsilon_2, \dots, \epsilon_p$ were independent we should, of course, have a set of p independent

regressions, one for each vector \mathbf{x} . We shall not assume that they are independent. By taking one variable z_1 as a dummy variable with value unity we may allow the component $\beta_{j1} z_1$ to represent means, and hence assume all the elements of ϵ to have a zero mean (cf. Exercise 19.1).

Our object is to estimate β and the dispersion matrix of ϵ which we shall write as σ . We write $\alpha = \sigma^{-1}$. (42.55) is, in different notation, the p -variate generalization of the general linear model of Chapters 19, 24, and subsequently. Here, however, we assume normality from the outset.

42.16 As in the univariate case we estimate β by maximizing the likelihood, which is given by

$$L = \frac{|\alpha|^{1/2n}}{(2\pi)^{1/2np}} \exp \left[-\frac{1}{2} \sum_{t=1}^n (\mathbf{x}_t - \beta \mathbf{z}_t)' \alpha (\mathbf{x}_t - \beta \mathbf{z}_t) \right], \quad (42.56)$$

where the suffix t is a sample label.

For the ML estimator of β_{jk} we find

$$\frac{\partial \log L}{\partial \beta_{jk}} = - \sum_{t=1}^n \sum_{s=1}^p \left\{ z_{kt} \alpha_{js} \left(x_{st} - \sum_{m=1}^p \hat{\beta}_{sm} z_{mt} \right) \right\} = 0. \quad (42.57)$$

Writing

$$u_{sk} = \sum_t z_{kt} x_{st} = \mathbf{x}_s' \mathbf{z}_k' \quad (42.58)$$

$$v_{sk} = \sum_t z_{kt} z_{st} = \mathbf{z}_s' \mathbf{z}_k', \quad (42.59)$$

we find from (42.57)

$$\sum_s \alpha_{js} \left(u_{sk} - \sum_{m=1}^p \hat{\beta}_{sm} v_{mk} \right) = 0$$

and as α is non-singular this gives us

$$u_{sk} - \sum_{m=1}^p \hat{\beta}_{sm} v_{mk} = 0 \quad (42.60)$$

which we may also write as

$$\hat{\beta} = \mathbf{u} \mathbf{v}^{-1}. \quad (42.61)$$

$p \times q \quad p \times q \quad q \times q$

For the estimator of α we have

$$\frac{\partial \log L}{\partial \alpha_{jk}} = \frac{1}{2} n \frac{A_{jk}}{|\alpha|} - \frac{1}{2} n \sum_{\mathbf{n}} S(x_j - \sum \beta_{jl} z_l)(x_k - \sum \beta_{km} z_m) = 0, \quad (42.62)$$

where A_{jk} is the cofactor of α_{jk} in $|\alpha|$. Bearing in mind that α is inverse to σ we find

$$\hat{\sigma}_{jk} = \frac{1}{n} S(x_j - \sum_l \hat{\beta}_{jl} z_l)(x_k - \sum_m \hat{\beta}_{km} z_m) \quad (42.63)$$

which we may also write

$$\hat{\sigma} = \frac{1}{n} (\mathbf{x} - \hat{\beta} \mathbf{z})(\mathbf{x} - \hat{\beta} \mathbf{z})'. \quad (42.64)$$

42.17 Since the z 's are fixed we find from (42.61)

$$\begin{aligned} E(\hat{\beta}) &= \{E(\mathbf{u})\} \mathbf{v}^{-1} = E(\mathbf{x}) \mathbf{z}' \mathbf{v}^{-1} \\ &= \beta \mathbf{z} \mathbf{z}' \mathbf{v}^{-1} = \beta. \end{aligned} \quad (42.65)$$

Writing temporarily \mathbf{V} for the inverse of \mathbf{v} we have from (42.61)

$$\hat{\beta}_{jk} = \sum_m u_{jm} V_{mk}. \quad (42.66)$$

From (42.55), multiplying by z_{lt} and summing over t , then by V_{lk} and summing over l , we find

$$\beta_{jk} = \sum_m u_{jm} V_{mk} - S \sum_t \epsilon_{jt} z_{lt} V_{lk}. \quad (42.67)$$

Hence

$$\hat{\beta}_{jk} - \beta_{jk} = S \sum_t \epsilon_{jt} z_{lt} V_{lk}. \quad (42.68)$$

Remembering that ϵ_{jt} and ϵ_{ku} are independent unless $t = u$ we find, with appropriate dummy suffixes for the summations,

$$\begin{aligned} E(\hat{\beta}_{jk} - \beta_{jk})(\hat{\beta}_{lm} - \beta_{lm}) &= E(S \sum_t \epsilon_{jt} z_{lt} V_{lk})(S \sum_t \epsilon_{lt} z_{\mu t} V_{\mu m}) \\ &= \sigma_{jl} S \sum_t \sum_{\lambda} \sum_{\mu} z_{\lambda t} z_{\mu t} V_{\lambda k} V_{\mu m} \\ &= \sigma_{jl} \sum_{\lambda} \sum_{\mu} v_{\lambda \mu} V_{\mu m} V_{\lambda k} \\ &= \sigma_{jl} \sum_{\lambda} \sigma_{\lambda m} V_{\lambda k} \\ &= \sigma_{jl} V_{mk} = \sigma_{jl} V_{km}. \end{aligned} \quad (42.69)$$

There are pq quantities β . The estimators $\hat{\beta}$ are distributed about mean β with dispersions given by (42.69). We may write equivalently

$$E(\hat{\beta}_j - \beta_j)'(\hat{\beta}_l - \beta_l) = \sigma_{jl} \mathbf{v}^{-1}. \quad (42.70)$$

As the β 's are linear functions of \mathbf{x} they are jointly normally distributed.

By putting $p = 1$ and transposing all matrices in (42.55), we return to the LS theory of Chapter 19, as the reader should verify for (42.61) and (42.70).

42.18 We may write

$$\begin{aligned} S(x_j - \sum_l \beta_{jl} z_l)(x_k - \sum_m \beta_{km} z_m) \\ = S(x_j - \sum_l \hat{\beta}_{jl} z_l)(x_k - \sum_m \hat{\beta}_{km} z_m) + S \sum_l (\hat{\beta}_{jl} - \beta_{jl}) z_l (\hat{\beta}_{km} - \beta_{km}) z_m, \end{aligned}$$

the cross-products vanishing,

$$= S(x_j - \sum_l \hat{\beta}_{jl} z_l)(x_k - \sum_m \hat{\beta}_{km} z_m) + \sum_l \sum_m (\hat{\beta}_{jl} - \beta_{jl})(\hat{\beta}_{km} - \beta_{km}) v_{lm}. \quad (42.71)$$

This is analogous to the univariate splitting of the sum of squares of errors into sum of squares of residuals (deviations from the estimated regression line) and a term due to the deviation of the estimated from the true parameter values—cf. (19.42), Vol. 2. It may be shown, by the argument used in reaching (42.69), that the last term on the right in (42.71) has an expectation of σ_{jk} . The first term on the right in (42.71) is a quadratic form in the x 's, which are multivariate normal. We do not, however, test one against the other, as in the univariate case, but the former against the whole.

42.19 To do this we require a theorem to the effect that the estimated dispersion $\hat{\sigma}$ of (42.64) is distributed in Wishart's form with $n - q$ instead of n . From (42.64), (42.59) and (42.60) we find the equivalent form

$$\hat{\sigma} = \frac{1}{n} \{ \mathbf{xx}' - \hat{\beta} \mathbf{v} \hat{\beta}' \}. \quad (42.72)$$

Our argument pursues the same line as the one used in 27.22, Vol. 2, in which we showed that, for normal variation, partial correlations have the same distribution as ordinary correlations, with lower degrees of freedom. Consider, in fact, a space of n dimensions. The $(q \times n)$ matrix \mathbf{z} determines a set of q vectors, say OP_{j1}, \dots, OP_{jq} in this space, themselves lying in a q -space.

Now

$$\epsilon = \mathbf{x} - \hat{\beta}\mathbf{z} + (\hat{\beta} - \beta)\mathbf{z}$$

and the two parts on the right are orthogonal in the n -space. For

$$\sum_{t=1}^n (x_{jt} - \sum_l \hat{\beta}_{jl} z_{lt}) \sum_m (\hat{\beta}_{jm} - \beta_{jm}) z_{mt} = \sum_m (u_{jm} - \sum_l \hat{\beta}_{jl} v_{lm}) (\hat{\beta}_{jm} - \beta_{jm})$$

and the first bracket vanishes in virtue of (42.60).

Thus the vectors $\mathbf{x} - \hat{\beta}\mathbf{z}$ are orthogonal to the q -space. Our original ϵ vectors are represented as the sum of two parts, one lying in the q -space and the other orthogonal to it. In our n -space, orthogonality implies zero correlation which implies independence for normal variables. Thus the two parts on the right in (42.71) are independent.

It follows that the system represented by $\mathbf{x} - \hat{\beta}\mathbf{z}$ has a Wishart distribution of dispersions, but with $n - q$ instead of n , the variation being orthogonal to a space of q dimensions.

42.20 We may now consider the testing of a hypothesis concerning regressions. Usually we require to know whether any of the β 's contribute significantly to the variation of x ; or equivalently, if we "extract" from x the variation due to a certain subset of (βz) 's, by the usual covariance technique, are the residuals significantly dependent on the remaining (βz) 's?

Suppose then that we take q β 's and test the hypothesis that a subset of $m \leq q$ are zero. On the hypothesis that they are not, we estimate the q β 's and σ and substitute in the likelihood (42.56). Now if we multiply (42.62) by $\hat{\alpha}_{jk}$ and sum over j, k , we find that the exponent in the likelihood reduces to a constant. Thus the likelihood of (42.56), apart from constants, reduces to $|\hat{\alpha}|^{\frac{1}{2}n}$. On the tested hypothesis with m β 's equal to zero, we find likewise that the likelihood reduces to some numerical multiple of, say, $|\hat{\alpha}_m|^{\frac{1}{2}n}$ where α_m is the inverse of the estimated dispersion matrix of (42.64), or equivalently of (42.72) with only $q - m$ β 's under estimate. Thus the likelihood ratio is

$$l = \frac{|\hat{\sigma}|^{\frac{1}{2}n}}{|\hat{\sigma}_m|^{\frac{1}{2}n}} \quad (42.73)$$

(42.73) is distributed as the ratio of two Wishart determinants based on $n - q$ and $n - (q - m)$ sample numbers. Moreover, by the same sort of argument as was used in 42.19, the vectors corresponding to the $q - m$ β 's supposed not to vanish are orthogonal in the q -space to the m β 's which are supposed to vanish, and hence the functions contributing to $\hat{\sigma}_m$ are an independent subset of those entering into $\hat{\sigma}$. Thus the criterion of (42.73) may be tested in the manner of the l criterion considered earlier in the chapter. The following example will illustrate the method.

Example 42.3 (Data from M. M. Barnard, 1935; Bartlett, 1947)

Miss Barnard had four series of Egyptian skulls, 91 Predynastic, 162 from the sixth to twelfth dynasties, 70 from the twelfth and thirteenth dynasties, and 75 from Ptolemaic dynasties—398 in all. On each skull four measurements were taken (in millimetres):

x_1 = maximum breadth

x_2 = basi-alveolar length

x_3 = nasal height

x_4 = basi-bregmatic height

The means of the series as given by Barnard are

	Series I $n_1 = 91$	Series II $n_2 = 162$	Series III $n_3 = 70$	Series IV $n_4 = 75$
x_1	133.582,418	134.265,432	134.371,429	135.306,664
x_2	98.307,692	96.462,963	95.857,143	95.040,000
x_3	50.835,165	51.148,148	50.100,000	52.093,333
x_4	133.000,000	134.882,716	133.642,857	131.466,667

The sums and squares of products within series (which have 394 degrees of freedom) are (42.74)

	x_1	x_2	x_3	x_4
x_1	9661.997,440	445.573,301	1130.623,900	2148.584,219
x_2		9073.115,207	1239.221,990	2255.812,722
x_3			3938.320,351	1271.051,662
x_4				8741.508,829

The similar sums for all observations together (397 d.fr.) are (42.75)

	x_1	x_2	x_3	x_4
x_1	9785.178,098	214.197,666	1217.929,248	2019.820,216
x_2		9559.460,890	1131.716,372	2381.126,040
x_3			4088.731,856	1133.473,898
x_4				9382.242,720

Finally, the sums of squares between classes (3 d.fr.) are

	x_1	x_2	x_3	x_4
x_1	123.180,628	-231.375,635	87.305,348	-128.763,994
x_2		486.345,863	-107.505,618	125.313,318
x_3			100.411,505	-137.580,764
x_4				640.733,891

(42.77)

We may first of all consider whether the data are homogeneous, in particular whether there are any significant differences between means of series. The appropriate criterion is the ratio l_2 of (42.17), namely the ratio of the determinants of (42.75) and (42.76) which is

$$l_2^{2/n} = \frac{2426 \cdot 898}{2954 \cdot 474} = 0.8214.$$

$-n \log l_2^{2/n}$, taking n as the number of degrees of freedom 397, is then 77.3 and the number of d.fr. for the approximate χ^2 test is $3 \times 4 = 12$. Thus we conclude that the data are not homogeneous even in the mean.

There are several questions we may wish to ask at this point. For example, from (42.76) it is clear that within classes there is a considerable correlation between the variables. Are the differences between the means attributable to influences from all four variables, or, for example, do x_3 and x_4 contribute to the differences only because of their correlation with x_1 and x_2 ? To answer this we determine the regressions of x_3 and x_4 on x_1 and x_2 , extract them from the total variation, and test the residual matrices. Thus we regard x_3 and x_4 as a matrix of dependent variables (the x 's of 42.16) and x_1, x_2 as z 's.

In our present case, the dispersion matrix of x_1, x_2 from (42.75) is

$$\frac{1}{394} \begin{bmatrix} 9661.977,040 & 445.573,301 \\ & 9073.115,207 \end{bmatrix}. \quad (42.78)$$

the inverse of which is

$$394 \times 10^{-4} \begin{bmatrix} 1.037,332 & -0.050,942 \\ & 1.104,659 \end{bmatrix}. \quad (42.79)$$

The variation due to regression of \mathbf{x} on \mathbf{z} is, from (42.72),

$$\hat{\beta} \mathbf{v} \hat{\beta}' = (\mathbf{u} \mathbf{v}^{-1}) (\mathbf{v}) (\mathbf{v}^{-1})' \mathbf{u}' = \mathbf{u} \mathbf{v}^{-1} \mathbf{u}'.$$

In our case \mathbf{x} refers to x_3 and x_4 , \mathbf{z} to x_1 and x_2 , so we find for this expression

$$\begin{aligned} 10^{-4} \begin{bmatrix} 1130.623,900 & 1239.221,990 \\ 2148.584,210 & 2255.812,722 \end{bmatrix} \begin{bmatrix} 1.037,332 & -0.050,942 \\ -0.050,942 & 1.104,659 \end{bmatrix} \\ = \begin{bmatrix} 1130.623,900 & 2148.584,210 \\ 1239.221,990 & 2255.812,722 \end{bmatrix} \\ = \begin{bmatrix} 287.967,620 & 534.238,796 \\ 534.238,796 & 991.621,041 \end{bmatrix}. \end{aligned} \quad (42.80)$$

Subtracting this from the matrix of x_3 and x_4 , we have as residual

$$\begin{bmatrix} 3650.353,731 & 736.815,866 \\ & 7749.887,788 \end{bmatrix}, \quad (42.81)$$

with $394 - 2 = 392$ d.fr.

Similarly, operating on (42.76) for the totals of product sums we find the residual

$$\begin{bmatrix} 3809.335,190 & 611.698,381 \\ & 8393.755,848 \end{bmatrix} \quad (42.82)$$

The question is whether the matrices (42.81) and (42.82) are significantly different. We can regard the latter as the residual in the regression of x_3, x_4 on x_1, x_2 plus a vector representing the mean; the former has had the mean abstracted in each class. The ratio (42.73) of their determinants is

$$l^{2/n} = \frac{0.277,469}{0.316,003} = 0.8781.$$

$-n \log l^{2/n}$, with $n = 392$, is then 51.39 and the appropriate number of degrees of freedom is $3 \times 2 = 6$. Homogeneity is therefore rejected. We conclude that x_3 and x_4 are relevant variables in the sense that the differences between means cannot be ascribed to x_1 and x_2 alone.

A further question considered by Miss Barnard was whether these variables might each have a linear regression on time. To investigate this we require a time variable, and the intervals between the four series were taken proportionately to 2, 1, 2. We may therefore conveniently take the values of t as $-5, -1, 1, 5$. On this basis

$$\begin{aligned} S(t-\bar{t})^2 &= 4307.663,32 \\ Sx_1(t-\bar{t}) &= 781.762,86 \\ Sx_2(t-\bar{t}) &= -1407.260,75 \\ Sx_3(t-\bar{t}) &= -410.101,94 \\ Sx_4(t-\bar{t}) &= -733.427,58 \end{aligned}$$

We are now examining the regression of each of the x 's on the extraneous variable time. The sums of squares and products due to regression (1 degree of freedom) are

	x_1	x_2	x_3	x_4
x_1	119.930,358	-234.810,812	68.428,625	-122.377,258
x_2		459.734,449	-133.975,163	-149.601,596
x_3			39.042,852	-69.824,358
x_4				124.874,099

(42.83)

Here, for example, the item in row 1 and column 2 is

$$\frac{Sx_1(t-\bar{t})Sx_2(t-\bar{t})}{S(t-\bar{t})^2} = \frac{(781.762,86)(-1407.260,75)}{4307.668,72} = -234.810,818.$$

The residual after removing the regression on time from the original matrix is given by subtracting (42.83) from (42.76), namely

	x_1	x_2	x_3	x_4
x_1	9665.247,740	449.008,478	1149.501,013	2142.197,474
x_2		9099.726,441	1265.691,535	2231.524,444
x_3			4049.689,004	1203.298,256
x_4				2957.368,621

(42.84)

with 396 d.fr.

We now test whether this residual is homogeneous, taking the variation within series as given by (42.79) against (42.84). The ratio of determinants is $|I^{2/n}| = 0.9031$. $-n \log |I^{2/n}|$ is 40 with $2 \times 4 = 8$ d.fr. We reject the hypothesis of homogeneity.

We conclude that if regression on time is linear, there are differences between the series which are not due to temporal effects.

42.21 The family of likelihood criteria which we have considered so far are ratios of determinants of dispersions of one kind or another and, algebraically speaking, are relatively simple. Other statistics which might be useful in testing are the value or values of latent roots of dispersion matrices. For example, the equality of two dispersion matrices **A** and **B** depends on the nearness to unity of the roots of $|A - \lambda B| = 0$. In a sense, then, tests of ratios of type $|A|/|B|$ can be carried out on latent roots. However, as we noted in 41.27, the exact distributions are not yet well tabulated, and although we may derive large-sample approximations, they are not so good as those for the likelihood criteria, which can be carried to any desired degree of accuracy by the methods of 42.11.

In fact, some of our likelihood ratio criteria are symmetric functions of the latent roots. For example, in $|A - \lambda B| = 0$ the product of all p roots is $|A|/|B|$, as is easily seen by writing $A = AB^{-1}B$. Cases where we are more interested in individual roots occur in the next chapter.

Example 42.4 (Foster and Rees, 1957. Data from Ashton, Lipton and Healy, 1957)

Data are given for two measurements ($p = 2$) on three groups of males: human, chimpanzee, and orang-utan. The measurements were on tooth length (x_1) and breadth (x_2) for the permanent upper second premolar, and were transformed to logarithms to stabilize variance.

The sums of products were as follows:

	d.fr.	Sums of products		
		x_1^2	x_1x_2	x_2^2
Between groups	2	0.544,941	0.525,765	0.509,075
Within groups	154	0.137,786	0.069,342	0.092,792
TOTAL	156	0.682,727	0.595,107	0.601,867

To test homogeneity we may consider the roots of (42.85)

$$|A - \lambda B| = 0$$

where **A** is the matrix between groups and **B** the total. The resulting equation is a quadratic with roots

$$\lambda_2 = 0.020,238, \quad \lambda_1 = 0.856,543.$$

We take the greater root as our criterion. The larger it is the more we suspect the hypothesis of homogeneity. From the Foster-Rees table (cf. 41.27(1)) the 99 per cent point of its distribution is found to be about 0.08. The observed value is highly in excess of this.

The group means were

	No. in group	\bar{x}_1	\bar{x}_2
Human	59	1.846	1.981
Chimpanzee	55	1.865	2.008
Orang-utan	43	1.986	2.119

Had we wanted to test the hypothesis that **A** and **B** are equal, without knowing which is larger, we should have had to test the smallness of the lesser root as well. In the general case this presents a theoretical difficulty, since the joint distribution of smallest and largest roots is not known. However, as might be expected, they tend to independence for large p .

Power of the tests

42.22 We have already remarked on the embarrassing profusion of parameters appearing in a multivariate situation. Our criteria of testing in the normal null case do not contain them because we estimate them all; but when we wish to specify alternatives in order to ascertain the power of a test we are in a position of some complexity.

For a test of a mean vector based on T^2 (cf. 41.16) the power can be ascertained from existing tables. In fact, if the parent vector is μ , the distribution of T^2 based on another vector μ_0 , namely

$$T^2 = n(\bar{\mathbf{x}} - \mu_0)' \mathbf{C}^{-1} (\bar{\mathbf{x}} - \mu_0), \quad (42.86)$$

has a non-central F distribution with p , $n-p$ degrees of freedom and non-centrality parameter $n(\mu - \mu_0)' \Upsilon^{-1} (\mu - \mu_0)$. We can then use non-central F , or one of the approximations to it—cf. 24.32—*provided that we can specify γ* . It may also be shown (Simaika 1941) that T^2 is uniformly most powerful in the class of tests whose power depends only on the non-centrality parameter—cf. 24.36 in the univariate case. Similar remarks apply in the two-sample case—cf. 41.17.

42.23 For further studies of distributions and tests in the multivariate case reference may be made to the books by T. W. Anderson (1958) and E. L. Lehmann (1959).

The problems associated with the Behrens-Fisher test for the difference of two means when variances are not equal have given rise to considerable controversy and a good deal of alleged paradox in multivariate extensions (see Mauldon (1955)). We noted in 21.15, Vol. 2, that the problem can, in fact, be solved by a method due to Scheffé which avoids these difficulties. This method has been generalized by Bennett (1951) to the multivariate case. See T. W. Anderson (1958, Section 5.6) and (1964), and Exercise 42.12.

For power functions see Seber (1964b), Darroch and Silvey (1963), Hogg (1961). Das Gupta *et al.* (1964) and T. W. Anderson and Das Gupta (1964a, b) obtained results on the monotonicity of the power functions of a number of tests of multivariate hypotheses.

Arnold (1964) has considered the distribution of T^2 under permutations, and Ito and Schull (1964) have discussed the robustness of the T_0^2 test, a generalization to several samples by Lawley (1939) and Hotelling (1951) of the two-sample T^2 test for the equality of mean-vectors. Mikhail (1965)—cf. also Ito (1962)—compares the power of T_0^2 , the

THE ADVANCED THEORY OF STATISTICS

LR test (42.17), and another test. The tests are asymptotically equivalent; in small samples, the LR test seems best. Schatzoff (1966) compares the tests using a different criterion, and concludes that there is little to choose between (42.17) and T_0^2 .

Posten and Bargmann (1964) give a method for obtaining the power of a LR test which is completely general for any hypothesis imposing one or two constraints.

42.24 Tamura (1966) gives asymptotic theory for distribution-free tests of the equality of location-parameter vectors in a set of otherwise identical continuous multivariate distributions.

EXERCISES

42.1 Show that for $p = 1$ the use of the criterion l_{H_2} of 42.6 becomes equivalent to an F -test.

42.2 By considering the maximization process over the various domains of the parameters, show that (42.18) is necessarily true.

42.3 Following Example 42.1, show that, for the criterion l_{H_1} , $-2\rho \log l$ is distributed approximately as χ^2 with $\frac{1}{2}(k-1)p(p+1)$ d.f. and

$$\rho = 1 - \left(\sum_{j=1}^p \frac{1}{r_j} - \frac{1}{r} \right) \frac{2p^2 + 3p - 1}{6(p+1)(k-1)}.$$

42.4 Show that $l^{2/n}$ of (42.44) can be represented as the product of independent variables y_{jk}

$$\prod_{j=2}^q \left\{ \prod_{k=1}^{p_j} y_{jk}^2 \right\}$$

where y_{jk} is a Beta-variable with parameters $\frac{1}{2}(n - \bar{p}_j - k)$, $\frac{1}{2}\bar{p}_j$

and

$$\bar{p}_j = \sum_{\alpha=1}^{j-1} p_{\alpha}.$$

42.5 Derive equation (42.41).

42.6 Derive equations (42.52)–(42.54).

42.7 A sample of n values is given from a single p -variate normal population. Consider the hypotheses

H : that means and dispersions are equal for each variate;

H_1 : that dispersions are equal regardless of means (i.e. all variances are equal and all covariances are equal);

H_2 : that given equality of dispersions, all means are equal.

Show that the likelihood ratio criteria are given by

$$l^{2/n} = \frac{|c_{jk}|}{(1-r_0)^{p-1} (s_0^2)^p \{1 + (p-1)r_0\}}$$

$$l_1^{2/n} = \frac{|c_{jk}|}{(1-r)^{p-1} (s^2)^p \{1 + (p-1)r\}}$$

$$(l_1 l_2)^{2/n(p-1)} = l,$$

where s^2 is the mean variance $\sum_{j=1}^p s_j^2/p$,

$$r = \sum_{j \neq k} c_{jk} / \{p(p-1)s_0^2\},$$

and s_0^2 , r_0 are the variance and correlation calculated from the pooled variables, e.g.

$$s_0^2 = \frac{1}{p} \sum_{j=1}^p \{s_j^2 + (\bar{x}_j - \bar{x})^2\}, \quad r_0 s_0^2 = r s^2 - \frac{1}{p(p-1)} \sum (\bar{x}_j - \bar{x})^2.$$

Show that $-2 \log l$, $-2 \log l_1$ and $-2 \log l_2$ are distributed approximately as χ^2 with $\frac{1}{2}p(p+3)-3$, $\frac{1}{2}p(p+1)-2$ and $p-1$ d.f. respectively.

(Wilks, 1946)

42.8 Use the distribution of (42.25) to confirm the conclusion of Example 42.4.

42.9 Verify the results of 42.12.

42.10 Derive T^2 as a likelihood ratio criterion in the form

$$l = \left(1 + \frac{T^2}{n-1}\right)^{-\frac{1}{2}n}$$

and derive its large-sample distribution in the null case.

42.11 Show further that in the general case, with a p -variate sample from $N(\mu, \gamma)$, and

$$T^2 = n(\bar{\mathbf{x}} - \mu_0)' \mathbf{c}^{-1} (\bar{\mathbf{x}} - \mu_0),$$

then $\frac{T^2}{n-1} \cdot \frac{n-p}{p}$ is distributed in the non-central F form with p , $n-p$ d.f. and non-centrality factor $n(\mu - \mu_0)' \gamma^{-1} (\mu - \mu_0) = \tau^2$, say.

42.12 $x_j^{(1)}$ ($j = 1, 2, \dots, n_1$), $x_j^{(2)}$ ($j = 1, 2, \dots, n_2$) are samples from two p -variate populations $N(\mu_1, \gamma_1)$, $N(\mu_2, \gamma_2)$. Define

$$y_j = x_j^{(1)} - \sqrt{\frac{n_1}{n_2}} x_j^{(2)} + \frac{1}{\sqrt{(n_1 n_2)}} \sum_{j=1}^{n_1} x_j^{(2)} - \frac{1}{n_2} \sum_{k=1}^{n_2} x_k^{(2)},$$

$$j = 1, 2, \dots, n_1, \quad k = 1, 2, \dots, n_2,$$

so that

$$\bar{\mathbf{y}} = \bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}.$$

Show that the covariance matrix of the y 's is given by

$$u_{\alpha\beta} = \delta_{\alpha\beta} \left(\gamma_1 + \frac{n_1}{n_2} \gamma_2 \right).$$

Defining \mathbf{w} by

$$(n_1 - 1) \mathbf{w} = \sum_{j=1}^{n_1} (\mathbf{y}_j - \bar{\mathbf{y}})(\mathbf{y}_j - \bar{\mathbf{y}})',$$

show that

$$T^2 = n_1 \bar{\mathbf{y}}' \mathbf{w}^{-1} \bar{\mathbf{y}}$$

is distributed as T^2 with $n_1 - 1$ d.f.

(Bennett, 1951)

42.13 Show that any test based on T^2 is invariant under a non-singular linear transformation of the variables with matrix say \mathbf{M} . By considering a transformation reducing the dispersion matrix \mathbf{c} to \mathbf{I} , show that the only invariant function involving only $\bar{\mathbf{x}}$ and \mathbf{c} is $\bar{\mathbf{x}}' \mathbf{c} \bar{\mathbf{x}}$.

42.14 Referring to Exercises 42.11 and 42.13, show that the distribution of T^2 may be written as a constant times

$$\exp(-\frac{1}{2}\tau^2) \sum_{j=0}^{\infty} \frac{(\frac{1}{2}\tau^2)^j}{j!} \frac{\{T^2/(n-1)\}^{\frac{1}{2}p+j} \Gamma(\frac{1}{2}n+j)}{\Gamma(\frac{1}{2}n+p) \{1+T^2/(n-1)\}^{\frac{1}{2}n+j}}.$$

Noting that the most powerful test using T^2 against $\tau^2 \neq 0$ is the ratio of this density to the value it takes when $\tau = 0$, show that the test is uniformly most powerful.

42.15 In the multivariate regression situation consider the partition of β into β_1, β_2 with q_1 and q_2 columns respectively. For testing the hypothesis $H: \beta_1 = \beta_1^*$ show that the $(2/n)$ th power of the likelihood ratio can be written

$$l = \frac{n\hat{\sigma}^2}{n\hat{\sigma}^2 + (\hat{\beta}_1 - \beta_1^*)' \mathbf{v}_{11.2} (\hat{\beta}_1 - \beta_1^*)},$$

where $\mathbf{v}_{11.2} = \mathbf{v}_{11} - \mathbf{v}_{12} \mathbf{v}_{22}^{-1} \mathbf{v}_{21}$ and the v 's are partitions of the matrix \mathbf{v} of (42.59) into q_1, q_2 rows and columns, viz. as

$$\begin{pmatrix} \mathbf{v}_{11} & \mathbf{v}_{12} \\ \mathbf{v}_{21} & \mathbf{v}_{22} \end{pmatrix}.$$

(T. W. Anderson, 1958)

42.16 Show that when the hypothesis H is true, the moments of l in the previous exercise are given by

$$E(l^t) = \prod_{j=1}^p \frac{\Gamma\{\frac{1}{2}(n - q + 1 - j) + t\} \Gamma\{\frac{1}{2}(n - q_2 + 1 - j)\}}{\Gamma\{\frac{1}{2}(n - q + 1 - j)\} \Gamma\{\frac{1}{2}(n - q_2 + 1 - j) + t\}}.$$

(T. W. Anderson, 1958)

CHAPTER 43

CANONICAL VARIABLES

43.1 Apart from problems of distributional mathematics, multivariate analysis suffers from one serious handicap in practical application: the difficulty of disentangling a complicated inter-relationship among the variables and of interpreting the results of the analysis. This leads us to attempt to reduce the *number* of variables, on the one hand; and to transform them to *independence*, on the other. The methods described in this chapter are motivated by one or both of these objectives.

Component analysis

43.2 As usual, we consider a row vector \mathbf{x}_j , $j = 1, 2, \dots, p$, representing a p -dimensional random variable and n observations on it, x_{jk} , $k = 1, 2, \dots, n$, resulting in a $p \times n$ matrix \mathbf{x} . It will often be convenient to measure each x_j about the mean of its n values, in which case the observed dispersion matrix \mathbf{c} is given by

$$\mathbf{c} = \frac{1}{n} \mathbf{x} \mathbf{x}' \quad (43.1)$$

We recall that if \mathbf{c} is of rank $m \leq p$ there are $p - m$ linear relations among the x 's. This implies that there is at least one linear transformation to new variables which are only m in number—our true dimensionality, so to speak, is m , fewer than p . The result derives from the fact, which is not difficult to prove, that the rank of a matrix multiplied by its transpose is the rank of the original matrix.

Example 43.1

Consider the $p \times p$ matrix

$$\begin{vmatrix} 1 & \rho & \rho & \cdot & \cdot & \cdot & \rho \\ \rho & 1 & \rho & \cdot & \cdot & \cdot & \rho \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho & \rho & \rho & \cdot & \cdot & \cdot & 1 \end{vmatrix}. \quad (43.2)$$

Add the rows and take out the common factor $1 + (p-1)\rho$. Subtract ρ times the resulting unit row from each other row. We then see that the determinant of the matrix is

$$(1-\rho)^{p-1} \{1 + (p-1)\rho\}. \quad (43.3)$$

Except in the special case $\rho = 1$ or $\rho = -1/(p-1)$ this cannot vanish. The rank of (43.2), accordingly, is p . Hence we cannot represent a set of equally correlated variables in fewer than p dimensions.

We may remark without proof (for which see Ledermann, 1937) that the number of independent conditions on a *symmetric* matrix for it to be of rank m is $\frac{1}{2}(p-m)(p-m-1)$.

43.3 We may represent the matrix \mathbf{x} geometrically in two different ways. We may set up a Euclidean space of p dimensions, one for each variable, and regard each sample set x_{jk} , $j = 1, 2, \dots, p$ as determining a point in it, so that our sample consists of a swarm of n points; or we may set up a space of n dimensions, one for each observation, and consider each variable in it, so that the variation is described by p vectors (lying in a p -dimensional space embedded in the n -dimensional space). In either kind of space, degeneration of the matrix \mathbf{x} to rank $m < p$ implies that the n sample points lie in a sub-space of m dimensions.

43.4 Consider a transformation to new variables ξ given by

$$\xi = \mathbf{a}\mathbf{x} \quad (43.4)$$

where \mathbf{a} is a matrix of coefficients. We confine our attention to linear transformations of this kind—non-linear situations are much more difficult to handle, and if they are suspected to exist an attempt should be made to linearize the data beforehand, for example, by a logarithmic transformation.

We shall, in fact, specialize \mathbf{a} to be orthogonal and call it \mathbf{l}' . Specifically,

$$\mathbf{l}'\mathbf{l} = \mathbf{l}\mathbf{l}' = \mathbf{I}. \quad (43.5)$$

We then have for the dispersions of the ξ 's, say $\mathbf{V}(\xi)$,

$$\mathbf{V}(\xi) = \mathbf{l}'\mathbf{c}\mathbf{l}. \quad (43.6)$$

It follows, of course, that

$$|\mathbf{V}(\xi)| = |\mathbf{c}|. \quad (43.7)$$

There are p^2 coefficients l . Equation (43.5) imposes $\frac{1}{2}p(p+1)$ conditions on them, $\frac{1}{2}p(p-1)$ for the off-diagonal products and p for the diagonals. There are thus $\frac{1}{2}p(p-1)$ degrees of freedom in the transformation. Geometrically, it is equivalent to a rotation in our p -space.

We may find one such transformation, at least, for which the ξ 's are uncorrelated, for this imposes $\frac{1}{2}p(p-1)$ conditions on them. If the resulting ξ 's have variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2$, represented by the diagonal matrix Σ , we have

$$\mathbf{l}'\mathbf{c}\mathbf{l} = \Sigma \quad (43.8)$$

and hence, in virtue of (43.5),

$$\mathbf{c} = \mathbf{l}\Sigma\mathbf{l}' \quad (43.9)$$

which is equivalent to

$$\mathbf{c}\mathbf{l} = \mathbf{l}\Sigma. \quad (43.10)$$

Considering the first row in the equation $\mathbf{c}\mathbf{l} = \mathbf{l}\Sigma$, we have

$$\sum_{k=1}^p c_{1k} l_{km} = \sigma_1^2 l_{1m}, \quad m = 1, 2, \dots, p \quad (43.11)$$

or

$$(\mathbf{c} - \sigma_1^2 \mathbf{I})\mathbf{l}_1 = \mathbf{0}, \quad (43.12)$$

where \mathbf{l}_1 is the first row of \mathbf{l} . Hence

$$|\mathbf{c} - \sigma_1^2 \mathbf{I}| = 0. \quad (43.13)$$

A similar equation is obeyed by the other values σ_r^2 . Hence the p values of σ^2 are

the latent roots of \mathbf{c} . The corresponding variables ξ are the latent vectors. We shall call them *principal components*.

43.5 In general there are p different latent roots λ of the matrix \mathbf{c} . We shall find it convenient to regard them as of diminishing magnitude, i.e. $\sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_p^2$. If and only if the last q are zero will the matrix \mathbf{c} become of rank $p-q$. The size of the latent roots thus gives us a test of the rank of the dispersion matrix. We may go further, and say that if σ_p^2 is small the variation is "nearly" in $p-1$ dimensions, and so on.

From the manner of derivation it is clear that the axes in our p -space of the first kind are orthogonal. But we have also transformed to variables which are uncorrelated. Thus the corresponding vectors in our p -space of the second kind (43.3) are also orthogonal. Evidently the transformation is unique, because \mathbf{c} has only one set of latent roots except in degenerate cases. Hence our transformation is the only one which simultaneously produces orthogonality in both the p -spaces.

43.6 Consider the variance of ξ_j

$$\text{var } \xi_j = \mathbf{l}_j' \mathbf{c} \mathbf{l}_j, \quad (43.14)$$

where \mathbf{l}_j is the j th column vector in \mathbf{I} . Suppose that we seek to maximize this, subject to the orthogonality condition

$$\mathbf{l}_j' \mathbf{l}_j = 1.$$

With a Lagrange multiplier λ we then have to maximize unconditionally

$$\mathbf{l}_j' \mathbf{c} \mathbf{l}_j - \lambda \mathbf{l}_j' \mathbf{l}_j \quad (43.15)$$

which, on differentiation by l_{jk} , $k = 1, 2, \dots, p$, leads to a set of equations summarized by

$$(\mathbf{c} - \lambda \mathbf{I}) \mathbf{l}_j = \mathbf{0}. \quad (43.16)$$

Comparison with (43.13) shows that the values of λ are again the latent roots. From (43.14) and (43.16) we find that the maximum value $\text{var } \xi_j$ is in fact the corresponding latent root. Our new variable ξ_1 thus has the property of possessing the greatest variance of any linear function of the x 's. ξ_2 will have the greatest variance among linear functions orthogonal to (uncorrelated with) ξ_1 ; and so on.

43.7 It is instructive to consider the same problem geometrically. Consider the n sample points in our first type of p -space, measured about their mean and standardized so as to have unit variance. Thus

$$\sum_{\alpha=1}^n x_{j\alpha} = 0. \quad (43.17)$$

$$\sum_{\alpha=1}^n x_{j\alpha}^2 = 1. \quad (43.18)$$

Take a line with current co-ordinates X and direction cosines u_i ,

$$\frac{X_1 - m_1}{u_1} = \frac{X_2 - m_2}{u_2} = \dots = \frac{X_p - m_p}{u_p}. \quad (43.19)$$

The sum of squares of distances of the n points from this line is given by S , say, where

$$S = \sum_{\alpha=1}^n \left[\sum_{j=1}^p (x_{j\alpha} - m_j)^2 - \left\{ \sum_{j=1}^p u_j (x_{j\alpha} - m_j) \right\}^2 \right]. \quad (43.20)$$

Let us evaluate m and u so that this is a minimum. The partial derivatives of (43.20) with respect to each m_j then vanish, giving

$$-S(x_{j\alpha} - m_j) + \sum_{\alpha} u_j \sum_{j=1}^p u_j (x_{j\alpha} - m_j) = 0. \quad (43.21)$$

In virtue of (43.17) this reduces to

$$\frac{m_j}{u_j} = \text{constant}, \quad j = 1, 2, \dots, p. \quad (43.22)$$

Hence the origin lies on the line (43.19) and we may take the m 's to be zero. This is what we might expect: the line goes through the centre of gravity of the points. Then, using (43.18) we have

$$S = p - \sum_{\alpha=1}^n \left(\sum_{j=1}^p u_j x_{j\alpha} \right)^2. \quad (43.23)$$

The u 's are subject to the orthogonalizing condition $\sum u_j^2 = 1$. We then have to minimize unconditionally

$$p - \sum_{\alpha=1}^n \left(\sum_{j=1}^p u_j x_{j\alpha} \right)^2 + \lambda \sum_{j=1}^p u_j^2. \quad (43.24)$$

Differentiation by u_k leads to

$$\sum_{\alpha=1}^n \sum_{j=1}^p x_{j\alpha} x_{k\alpha} u_j - \lambda u_k = 0 \quad (43.25)$$

or

$$\sum_{j=1}^p r_{jk} u_j - \lambda u_k = 0. \quad (43.26)$$

The elimination of the u 's leads us back to

$$|\mathbf{r} - \lambda \mathbf{I}| = 0. \quad (43.27)$$

Thus the appropriate λ is a latent root of the correlation matrix. If we had not standardized by reducing the initial variation to unit variance we should have arrived at the latent roots of the dispersion, not the correlation, matrix.

Moreover, from (43.23) and (43.25) we find

$$S = p - \lambda. \quad (43.28)$$

It follows that the latent root which gives the minimum value to S is the largest latent root. Our line corresponds to ξ_1 and is such that the sum of squares of distances of sample points from it is a minimum.

We can now project all our points on to a hyperplane perpendicular to the line (43.19) and repeat the process by finding a line in that hyperplane such that the sum of squares of distances from the projected points is a minimum. Our line in the $(p-1)$ -space will be given by the second largest latent root of (43.27). This is not

immediately obvious. However, we saw in 43.5 that the second latent vector is orthogonal to the first and hence lies in the $(p-1)$ -space; and that it was derived by maximizing a variance, which is equivalent to minimizing the sum of squares of distances from the line.

43.8 The following points may be briefly noted:

- (1) The latent roots of a dispersion matrix are all real and non-negative. This stems from the fact that \mathbf{c} is non-negative definite. A formal proof will be found in most textbooks on matrix theory. See, however, the warning in 43.36.
- (2) In general the latent roots are unequal, but some or all of them may be equal in particular cases. Where equality exists, there is no criterion based on variance size to pick out any one (among the group with equal latent roots) as having priority. Any orthogonal set will do.
- (3) The sum of the latent roots, from (43.13), is the sum of the terms in the diagonal of the dispersion matrix, namely its trace. Likewise the product of the latent roots is the determinant of the dispersion matrix.
- (4) If \mathbf{A} and \mathbf{B} are both non-degenerate dispersion matrices the latent roots of $|\mathbf{A} - \lambda\mathbf{B}| = 0$ are the same as those of $|\mathbf{B}^{-1}\mathbf{A} - \lambda\mathbf{I}| = 0$. In particular, if $\mathbf{A} = \mathbf{I}$ we see that the latent roots of the inverse are the reciprocals of the latent roots of the matrix.

43.9 The question of standardization requires more attention. It has been customary, especially in psychological work, to standardize the dispersion matrix by dividing by appropriate (sample) standard deviations and hence to reduce it to the correlation matrix. In such a case the sum of the variances of the ξ 's is equal to the dimension number p . In effect, the procedure reduces all the variables to equal importance as measured by scale.

However, the latent roots and latent vectors are not invariant under changes of scale. In the geometrical representation of 43.7 perpendicular distances are no longer perpendicular. Thus, in general, we get different results according to whether a scale is initially imposed on the system or not. The point is illustrated in Example 43.4 later. Whether standardization is desirable is, in the ultimate analysis, to be decided on non-statistical grounds. From the statistical viewpoint it is a nuisance, especially in sampling investigations, because it complicates the distributional theory.

43.10 The actual solution of equation (43.13) by desk-machine is a rather tedious matter. For details of the iterative process involved see Kendall (1961b). The advent of the electronic machine has altered the arithmetical situation completely, and most machines are programmed to handle quite large matrices and print out the appropriate latent roots and latent vectors. We shall therefore not allot space to the problem of computation.

Example 43.2 (Lawley and Maxwell, 1963)

Five psychological tests were carried out on 123 individuals. The correlations between scores on the tests were as follows:

	1	2	Test 3	4	5
1.	1.	0.488	-0.137	0.205	-0.178
		1.	0.031	0.180	-0.304
			1.	0.161	0.372
				1.	-0.013
					1.

A principal component analysis gives the following latent roots and vectors:

Latent roots	1	2	Vectors 3	4	5
1.75714	.55550	.56470	-.27000	.23572	-.49403
1.33070	-.18568	-.24745	-.66199	-.55654	-.39538
0.78086	.21597	.43969	.32041	-.78704	.19478
0.70916	.64078	-.30765	-.39839	-.01481	.57950
0.42214	.44688	-.57611	.47696	-.12266	-.47523

The matrix l is given by the five columns on the right. For example

$$\xi_1 = .55550x_1 + .56470x_2 - .27000x_3 + .23572x_4 - .49403x_5$$

and, reading downwards,

$$x_1 = .55550\xi_1 - .18568\xi_2 + .21597\xi_3 + .64078\xi_4 + .44688\xi_5.$$

The original data, of course, hardly bear five-figure accuracy in these results, but it is convenient to retain them for checking purposes.

In psychological work it is customary to express these coefficients in a modified form. Instead of the variables ξ we introduce

$$\zeta_j = \xi_j / \sqrt{\lambda_j}$$

so that the ζ 's have unit variance. The matrix of coefficients of the x 's in terms of ζ 's is then

x 's	1	2	ζ 's 3	4	5
1	.73635	-.21419	.19081	.53961	.29035
2	.74855	-.28545	.38853	-.25908	-.37432
3	-.35790	-.76364	.28313	-.33549	.30989
4	.31247	-.64200	.69548	-.01247	-.07969
5	-.65488	-.45609	.17212	.48801	-.30877

Thus, for example,

$$x_1 = .73635\zeta_1 - .21419\zeta_2 + .19081\zeta_3 + .53961\zeta_4 + .29035\zeta_5. \quad (43.35)$$

These coefficients are known as factor loadings, the ζ 's being regarded as factors structural to the situation and the coefficients as weights with which they appear in the different variables.

We will deal with questions of testing and estimation presently. Taking the data as they stand, we see that the variables ξ (which by definition are uncorrelated) account, in turn, for 35, 27, 16, 14 and 8 per cent of the variance. These numbers are obtained by dividing the latent roots by p , in this instance 5. We might, as an approximation, be willing to omit the last variate, in which case the first four contribute 92 per cent of the variance; or even the last two variates, the other three contributing 78 per cent. But we still require measurements on all five x 's to calculate these three ζ 's.

43.11 We have noted that the latent roots are uniquely determined by the dispersion matrix. From (43.16), whether relating to sample or parent, it is clear that l_j is also uniquely determined, except perhaps for a change of sign, which we can always determine by taking l_{j1} as positive. Thus there is a one-to-one relation between the latent roots and latent vectors, and the dispersion matrix and the mean vectors. Since the sample values of the latter are the ML estimators of the corresponding parent values, the sample values of the latent roots or vectors are ML estimators of the parent values in normal variation.

The problem of bias has been considered by Dempster (1966). Exact expressions are complicated.

Testing of latent roots

43.12 An exact theory for testing latent roots is difficult to attain, for several reasons. Distributions are complicated; standardization procedures, as already noted, further complicate the issue; and we may be interested in the special cases where the latent vectors are indeterminate in the sense that a group of latent roots may be equal.

Let us be clear about the kind of hypothesis which we wish to test. The first is whether the sort of transformation which we have been discussing is worth while at all. This is equivalent to asking whether the latent roots are different from one another. If they are not, the original x 's are just as good as the ξ 's for purposes of representation. To put it another way, are the x 's independent?

We arrived at a test of this hypothesis in 42.12. The criterion is then the correlation determinant, $-n$ times the logarithm of which is distributed approximately in the χ^2 form with $\frac{1}{2}p(p-1)$ degrees of freedom. More accurately,

$$-n\left(1 - \frac{2p+11}{6n}\right) \log |r|$$

is distributed as χ^2 .

Example 43.3

In the data of Example 43.2 the value of the correlation determinant is (being the product of the latent roots) 0.54659. The value of n is 123. Thus, approximately, $-123 \log 0.54659 = 74.3$ is a χ^2 value with 10 d.f. This is extremely unfavourable to the hypothesis.

For the more accurate approximation we have that

$$-\left(n - \frac{2p+11}{6}\right) \log |r| = 72.2$$

is a χ^2 also with 10 d.f. The conclusion is unaffected.

This test is one of independence. If we wish to test both independence and equality of variance we use the sphericity test of 42.13 applied to the dispersion matrix \mathbf{c} . The criterion is then

$$-n\{\log |\mathbf{c}| - p \log (\text{trace } \mathbf{c}/p)\} \quad (43.36)$$

with $\frac{1}{2}p(p+1) - 1$ d.f. For the more refined test we replace n , using (42.54), by

$$n - \frac{2p^2 + p + 2}{6p(n-1)}. \quad (43.37)$$

43.13 This kind of test reveals whether there is any point in transforming to canonical variables ξ .

In one sense no test is required for non-vanishing latent roots. Any value which is greater than zero cannot have originated from a population in which the corresponding parent value was truly zero; for, if it had, the parent variation would lie in a sub-space and no sample point could arise from outside that space. This will not necessarily be so if the variate values are subject to errors of observation and measurement, but this case must be deferred for consideration until we examine factor analysis later.

43.14 We may, however, legitimately ask this kind of question: suppose that certain latent roots are large and account for most of the variance; do the remaining values differ significantly among themselves, or could they have arisen from a complex in which the corresponding variables are effectively spherical or at least uncorrelated? To put it another way, are the remaining λ 's and their associated latent vectors *distinguishable*?

Bartlett (1954 and earlier papers) proposed on somewhat heuristic grounds to test such a hypothesis by an approximate χ^2 test. Suppose that we have decided to retain the first k latent roots and wish to test whether the remaining $p-k$ are equal to some unknown value. We assume that the sampling errors are small enough, compared with the differences among $\lambda_1, \dots, \lambda_k$, for us to be able to set up an almost certain correspondence between the parent λ_j and the sample λ_j for $j = 1, 2, \dots, k$. Since the dispersion determinant is the product of the latent roots it seems reasonable to test that determinant against the one which would be reached if the last $p-k$ roots were all equal. In the particular case of a correlation determinant this value is

$$\lambda_1 \lambda_2 \dots \lambda_k \left\{ \frac{p - \lambda_1 - \lambda_2 \dots \lambda_k}{p-k} \right\}^{p-k}. \quad (43.38)$$

The criterion proposed is therefore the ratio of the two determinants, namely

$$(\lambda_{k+1}\lambda_{k+2}\dots\lambda_p)^{-1} \left\{ \frac{\lambda_{k+1} + \lambda_{k+2} + \dots + \lambda_p}{p-k} \right\}^{p-k} \quad (43.39)$$

where the λ 's are sample values. This may be regarded as the $(p-k)$ th power of the ratio of the arithmetic to the geometric mean of $\lambda_{k+1}, \dots, \lambda_p$.

The proposal is that the logarithm of this quantity, multiplied by a factor involving n , should be tested in the χ^2 distribution with

$$\frac{1}{2}(p-k-1)(p-k+2) \text{ d.fr.} \quad (43.40)$$

Lawley (1956a) has shown that if the multiplier is taken as

$$n-k-1 - \frac{1}{6} \left\{ \frac{2(p-k)^2 + p-k+2}{p-k} \right\} + \lambda^2 \sum_{j=1}^k \frac{1}{(\lambda_j - \lambda)^2}, \quad (43.41)$$

λ being estimated as the mean of $\lambda_{k+1}, \dots, \lambda_p$, the criterion has the correct moments for a χ^2 distribution to $O(n^{-3})$. If $\lambda_1, \dots, \lambda_k$ are large compared to λ the last term in (43.41) could be omitted.

43.15 Strictly speaking, these results apply to the dispersion matrix with units in the diagonals. Application to a correlation matrix is impaired by the fact that we standardize using the sample variance. It appears that in this case the criterion does not follow a χ^2 distribution. However, a rough test may be obtained, *faute de mieux*, by using the results of 43.14 as if they applied to a correlation matrix.

In this connexion, consider again the data of Examples 43.2 and 43.3. Suppose we decide that the two largest roots are different enough to justify a supposition that they are distinct among themselves and also distinct from smaller values.

The product of the remaining three roots is 0.23376 and their mean is 0.63739. The multiplier of (43.41), neglecting the last two terms, is 120, and the criterion becomes

$$120 \{ 3 \log 0.63739 - \log 0.23376 \} = 12.3.$$

From (43.40) the number of degrees of freedom is 5, and the observed value exceeds the 5 per cent point but not the 1 per cent. We suspect that the last three roots are genuinely unequal.

Large-sample results for latent roots

43.16 We can make further progress by considering asymptotic theory, namely standard errors and covariances when the parent latent roots are all distinct. The results were first obtained by Girshick (1939).

It is indifferent, to our order of approximation, whether we write our formulae in terms of parent values or of sample values. We will use sample values. We then have the following relations:

$$\sum_{\alpha} l_{j\alpha} l_{\alpha k} = \delta_{jk} \quad (43.42)$$

$$\sum_{\alpha} c_{j\alpha} l_{\alpha k} = \lambda_j l_{jk}, \quad (43.43)$$

and, using the Kronecker delta as before, we derive from (43.43)

$$\sum_{\alpha} c_{j\alpha} l_{\alpha k} l_{km} = \lambda_j \delta_{jm}. \quad (43.44)$$

From (43.43) we then have

$$\sum_{\alpha} c_{j\alpha} dl_{\alpha k} + \sum_{\alpha} dc_{j\alpha} l_{jk} = \lambda_j dl_{jk} + d\lambda_j l_{jk}. \quad (43.45)$$

Without loss of generality we may now suppose the axes rotated to the ξ -axes, in which case $c_{jj} = \lambda_j$ and $c_{jk} = 0, j \neq k$. Then the first terms on each side of (43.45) cancel and we find

$$dc_{jj} = d\lambda_j. \quad (43.46)$$

Thus

which, by use of (41.98), gives us for the normal case

$$\text{cov}(\lambda_j, \lambda_k) = \frac{2\lambda_j^2}{n} \delta_{jk}. \quad (43.47)$$

Hence λ_j, λ_k are uncorrelated for $j \neq k$ and

$$\text{var } \lambda_j = \frac{2\lambda_j^2}{n}. \quad (43.48)$$

To our approximation this entails that

$$\text{var}(\log \lambda_j) = 2/n, \quad (43.49)$$

a convenient form since the variance does not depend on the parameters λ .

43.17 Once again, the results for a correlation matrix, as distinct from a dispersion matrix, are much more complicated. We quote the results from Girshick (1939):

$$\text{cov}(\lambda_j, \lambda_k) = \frac{2}{n} \left\{ \sum_{\alpha, \beta} l_{\alpha j}^2 l_{\beta k}^2 r_{\alpha\beta}^2 - (\lambda_j + \lambda_k) \sum_{\alpha} l_{\alpha j}^2 l_{\alpha k}^2 \right\}, \quad (43.50)$$

$$\text{var } \lambda_j = \frac{2}{n} \left(\lambda_j^2 + \sum_{\alpha, \beta} l_{\alpha j}^2 l_{\beta j}^2 r_{\alpha\beta}^2 - 2\lambda_j \sum_{\alpha} l_{\alpha j}^4 \right) \quad (43.51)$$

where $r_{\alpha\beta}$ typifies correlations.

43.18 The same method may be used to derive variances and covariances for the coefficients of the latent vectors. For what they are worth, we quote the results; but it must be remembered that in practice we should rarely wish to test an individual direction cosine.

$$\text{cov}(l_{jk}, l_{mt}) = \frac{-\lambda_k \lambda_l l_{jt} l_{mk}}{n(\lambda_k - \lambda_l)^2}, \quad k \neq t, \quad (43.52)$$

$$\text{var } l_{jk} = \frac{l_{jk}^2}{4n} + \frac{\lambda_k^2}{n} \left(\frac{l_{j1}^2}{(\lambda_k - \lambda_1)^2} + \dots + \frac{l_{jk}^2}{4\lambda_k^2} + \dots + \frac{l_{jp}^2}{(\lambda_k - \lambda_p)^2} \right), \quad (43.53)$$

while $\text{cov}(l_{jk}, l_{mk})$ is given by (43.53) with each l_{js}^2 replaced by $l_{js} l_{ms}$.

For some recent work see T. W. Anderson (1963), who proves the asymptotic normality of the distribution of latent roots and vectors and deals with the case where some of the parent roots are equal.

43.19 As a statistical tool, principal component analysis is, perhaps, best regarded as an exploratory instrument to enable us to see what is the effective number of dimensions, or how dominant are certain linear combinations of the variables. There is

one case in which the use of the analysis has been pushed further, perhaps beyond allowable limits. Suppose that the largest latent root is dominant, accounting for, say, 70 or 80 per cent of the variance. It may be a rather Procrustean procedure to neglect the remainder and to force the whole variation, so to speak, in the direction of the first latent vector. But there are occasions when we are willing to do this; for example, if the x 's are values of business activity indices of one kind or another, bank deposits, freight loadings, imports and so on, we may be willing to allow the first principal component to determine a *single* number expressing the general intensity of business activity. The values of ξ_1 then become a weighted *index number* of the constituent values of x . Whether this index remains a pure artefact, or whether it corresponds to some "real" intensity of business activity, is a matter of interpretation to be decided in the light of our knowledge about the economic structure of the system under study.

Kendall (1961b) has shown how a fair approximation to the ordering of a set by the first principal component can be attained by ranking methods. Little else seems to be known about distribution-free methods in the field of canonical analysis.

Example 43.4 (Craddock, 1965 with some supplementary information kindly supplied by him in correspondence)

Manley (1953 and later) has constructed a remarkably long series of monthly temperatures for Central England from 1680 to 1963. The data are in degrees Fahrenheit to the nearest tenth of a degree. The year, for the purpose of the analysis, was taken to run from November through the following October.

Each year was taken as a 12-dimensional quantity, one value for each month. Thus no scaling problem arises. The variate values were measured from the mean of the whole series, not the individual monthly means. This leaves in the picture the variation of temperature over the year, and we shall not be surprised to find annual variation in a dominant position.

There were thus 283 sets of monthly mean temperatures running from November 1680 to October 1963.

In the treatment earlier in this chapter we have assumed the values of each component of x to be measured from the mean of that component. If we measure from some other value, our product-sums divided by n are no longer covariances but second-order moments. The analysis remains valid, but we must expect one component, probably the first, to correspond to an axis from the alternative mean through the sample mean.

The product-moment matrix is shown in Table 43.1, overleaf.

Table 43.1—Product-moment matrix of Central England temperatures, November 1680–October 1963:
data treated as departures from general mean of 48.45°F

Month	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.
Nov.	39.365	54.697	64.048	54.976	41.495	13.035	–20.322	–53.065	–70.060	–66.125	–42.323	–3.296
Dec.	54.697	93.814	104.526	89.371	66.342	21.303	–33.432	–84.421	–110.803	–105.328	–67.910	–5.906
Jan.	64.048	104.526	133.779	108.865	79.234	25.020	–40.215	–101.108	–131.932	–125.252	–80.368	–7.037
Feb.	54.976	89.371	108.865	102.280	71.219	22.593	–34.478	–88.008	–114.819	–108.928	–70.748	–5.630
Mar.	41.495	66.342	79.234	71.219	57.236	17.522	–25.164	–65.221	–85.348	–80.824	–51.806	–3.615
Apr.	13.095	21.303	25.020	22.593	17.522	10.216	–7.560	–19.716	–25.729	–24.331	–15.728	–0.756
May	–20.322	–33.432	–40.215	–34.478	–25.164	–7.560	18.137	35.331	45.067	42.970	27.787	3.340
June	–53.065	–84.421	–101.108	–88.008	–65.221	–19.716	35.331	90.261	114.083	107.854	69.624	7.044
July	–70.060	–110.803	–131.932	–114.819	–85.348	–25.729	45.067	114.083	151.505	141.129	90.662	9.108
Aug.	–66.125	–105.328	–125.252	–108.928	–80.824	–24.331	42.970	107.854	141.129	136.237	86.696	9.030
Sept.	–42.323	–67.910	–80.368	–70.748	–51.806	–15.728	27.787	69.624	90.662	86.696	59.000	6.237
Oct.	–3.296	–5.905	–7.037	–5.630	–3.615	–0.756	3.240	7.044	9.108	9.030	6.237	5.507
General mean	48.45	48.45	48.45	48.45	48.45	48.45	48.45	48.45	48.45	48.45	48.45	48.45
Variances	6.27	9.69	11.57	10.11	7.57	3.20	4.26	9.50	12.31	11.67	7.68	2.35

The first ten latent roots of this matrix are as follows:

Latent root number	Value as percentage of variance
1	92.38
2	2.05
3	1.12
4	0.98
5	0.67
6	0.58
7	0.49
8	0.45
9	0.41
10	0.36
Sum of first 10	99.47

The amount of variation accounted for by the first latent vector is unusually high, but there is, of course, a reason—the major variation is a seasonal movement. The coefficients of the first four latent vectors are given in Table 43.3 later. Plotted against the monthly means given in Table 43.2, they are seen to pursue an almost identical pattern of seasonal movement.

In psychological or economic work we should hardly bother to consider the other latent roots. However, the interest of the present example is that sufficient knowledge is available of the physical system which generates the data to enable some attempt at interpretation. Craddock, to whose paper reference may be made for details, identifies the second component with climatic changes in the annual mean temperature, and the third and fourth with patterns of variation in winter temperature.

Table 43.2 gives the covariances, moments being measured about the monthly means.

The first four latent roots of this matrix are:

Latent root number	Value as percentage of variance
1	27.50
2	15.20
3	10.84
4	10.78
Sum of first 4	64.32

The picture of residual variation, after the abstraction of seasonal components, is now much less clear. From the coefficients of the latent vectors, which are given in Table 43.3, it appears that the first (whose coefficients are all positive) represents movement of a secular kind; the second and third indicate a harmonic movement over the year and, as in the analysis of Table 43.1, seem to represent a type of variation in winter temperature.

Table 43.2—Covariance matrix of Central England temperatures, November 1680 - October 1963:
data treated as departures from the individual monthly means

Month	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.
Nov.	6.051	1.722	0.887	-0.129	0.595	0.334	0.876	0.616	-0.021	0.308	0.424	0.782
Dec.	1.722	9.574	4.090	1.745	1.305	1.106	0.276	0.909	0.570	0.311	0.066	0.580
Jan.	0.887	4.090	14.032	4.392	1.693	0.940	-0.026	0.629	0.855	0.698	0.677	0.695
Feb.	-0.129	1.745	4.392	11.132	3.568	1.584	0.585	0.752	1.031	0.957	-0.039	1.116
Mar.	0.595	1.305	1.693	3.568	7.025	1.929	0.860	0.658	0.637	0.734	0.675	1.392
Apr.	0.334	1.106	0.940	1.584	1.929	5.374	0.521	0.742	0.973	0.997	0.569	0.798
May	0.876	0.276	-0.026	0.585	0.860	0.521	4.648	1.186	0.501	0.699	0.587	0.645
June	0.616	0.909	0.629	0.752	0.658	0.973	1.186	3.827	1.269	0.847	0.768	0.475
July	-0.021	0.570	0.855	1.031	0.637	0.973	0.501	1.269	4.259	1.464	0.792	0.534
Aug.	0.308	0.371	0.698	0.957	0.734	0.997	0.699	0.847	1.464	3.763	1.452	0.898
Sept.	0.424	0.066	0.677	-0.039	0.675	0.569	0.587	0.768	0.792	1.452	4.148	1.004
Oct.	0.782	0.580	0.695	1.116	1.392	0.798	0.645	0.475	0.534	0.898	1.004	5.008
Monthly means	42.68	39.27	37.51	38.91	41.37	46.25	52.13	57.75	60.59	59.96	55.86	49.16
Variances	2.46	3.09	3.75	3.34	2.65	2.32	2.16	1.96	2.06	1.94	2.04	2.24

It is interesting to consider what happens to this analysis if we standardize by reducing the covariances of Table 43.2 to correlations. The corresponding figures are:

Latent root number	Value as percentage of variance
1	22.54
2	11.51
3	10.29
4	8.93
Sum of first 4	53.27

Although the differences are not so very large, they are appreciable. Even in this case, then, when all the p components of the vector are measured in the same units, standardization makes a difference. We should expect greater differences in cases where the components are measured in units of different range.

Canonical correlations

43.20 The transformation of a set of variables x to a canonical set ξ is effectively the reduction of a quadratic form to a sum of squares by linear transformation. We now turn to consider the general theory of the relations between two sets of variates x_1, \dots, x_p and x_{p+1}, \dots, x_{p+q} , where we suppose that $p \leq q$. Following Hotelling (1936) we shall show that in general there can be found linear transformations to variates ξ_1, \dots, ξ_p ; $\xi_{p+1}, \dots, \xi_{p+q}$, such that

- All the ξ 's have unit variance and zero mean;
- any ξ in the p -group is uncorrelated with the other ξ 's in that group;
- any ξ in the q -group is uncorrelated with the other ξ 's in that group;
- the correlation between any ξ in the p -group and any ξ in the q -group is zero except for p correlations $\rho_1, \rho_2, \dots, \rho_p$, which may be taken to be the correlations between ξ_1 and ξ_{p+1} , ξ_2 and ξ_{p+2} , \dots , ξ_p and ξ_{2p} .

The variates ξ are then said to be in canonical form and the ρ 's are called canonical correlations. We have already discussed canonical correlation in the context of the analysis of categorized data in 33.44-9, Vol. 2.

In the case of a single set of variables we were able to ensure that the ξ 's in turn accounted for as much as possible of the total variation. This is no longer possible here. The optimization is concerned with the reduction in the intercorrelations to a minimal set.

We will suppose that our variables x have zero means and dispersions typified by γ_{jk} . Those dispersions in the p -group we denote by Greek suffixes, $\gamma_{\alpha\beta}$, and those in the q -group by Roman suffixes, γ_{jk} . For the covariance of a p -variate and a q -variate we write one Greek and one Roman suffix: $\gamma_{\alpha j}$.

To simplify the notation we will omit suffixes referring to sample labels. Indeed, we can go further and omit other suffixes identifying ξ , η and the corresponding

Table 43.3—Coefficients of latent vectors representing Central England temperature variation, November 1680-October 1963

Covariance matrix—data treated as departures from general mean		Principal component vectors representing Central England temperature variation, November 1680-October 1963											
Latent vector	Contribution (percentage)	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.
1	92.38	-.201	-.323	-.387	-.337	-.248	-.077	.128	.322	.421	.399	.258	.024
2	2.05	.037	.234	.473	.397	.239	.203	.177	.297	.361	.343	.244	.193
3	1.12	-.139	-.399	-.515	.560	.430	.180	.063	-.057	-.044	-.014	-.061	.110
4	0.97	.498	.498	-.485	-.268	.238	.213	.205	.099	-.046	.003	.054	.194
Covariance matrix—data treated as departures from monthly means													
1	27.50	.108	.377	.640	.501	.293	.179	.072	.110	.121	.113	.074	.132
2	15.20	.038	.341	.673	-.374	-.415	-.223	-.157	-.047	-.017	-.086	-.041	-.178
3	10.84	-.399	-.423	.449	.226	-.219	-.271	-.246	-.222	-.279	-.165	-.159	-.218
4	10.78	-.377	.426	-.261	.489	-.003	-.037	-.235	-.109	-.114	-.211	-.384	-.310
Correlation matrix—data treated as departures from monthly means													
1	22.54	.169	.244	.262	.310	.343	.317	.237	.316	.315	.350	.271	.279
2	11.51	-.037	.252	.286	.498	.380	.136	-.204	-.260	-.260	-.331	-.395	-.020
3	10.29	.482	.471	.450	-.157	-.341	-.338	.044	.140	.138	.153	-.027	-.152
4	8.93	.347	.015	-.467	-.033	.221	.090	.560	.170	-.305	-.288	-.275	.091

coefficients in the transformation. Consider now a particular pair of variables, one from each group, given by

$$\xi = \sum_{\alpha} l_{\alpha} x_{\alpha}, \quad \alpha = 1, 2, \dots, p; \quad (43.54)$$

$$\eta = \sum_a m_a x_a, \quad a = p+1, p+2, \dots, p+q. \quad (43.55)$$

We take them to have unit variances and hence

$$\sum_{\alpha, \beta} l_{\alpha} l_{\beta} \gamma_{\alpha\beta} = 1, \quad (43.56)$$

$$\sum_{a, b} m_a m_b \gamma_{ab} = 1. \quad (43.57)$$

We now seek the condition that their correlation R is stationary for variations in the coefficients l and m , namely that

$$R = \sum_{\alpha, a} l_{\alpha} m_a \gamma_{\alpha a} \quad (43.58)$$

is stationary. Taking two undetermined multipliers $\frac{1}{2}\lambda$ and $\frac{1}{2}\mu$, we then have to find an unconditioned stationary value of

$$\sum l_{\alpha} m_a \gamma_{\alpha a} - \frac{1}{2}\lambda \sum l_{\alpha} l_{\beta} \gamma_{\alpha\beta} - \frac{1}{2}\mu \sum m_a m_b \gamma_{ab}. \quad (43.59)$$

On differentiation this leads to

$$\begin{aligned} \sum_{\alpha} l_{\alpha} \gamma_{\alpha a} - \mu \sum_b m_b \gamma_{ab} &= 0, \\ \sum_a m_a \gamma_{\alpha a} - \lambda \sum_{\beta} l_{\beta} \gamma_{\alpha\beta} &= 0. \end{aligned} \quad (43.60)$$

Multiplying the first equation by m_a and summing, and the second by l_{α} and summing, we find, in virtue of (43.56)–(43.58),

$$R = \lambda = \mu. \quad (43.61)$$

Equations (43.60) are then solvable for l and m if their determinant vanishes. Writing λ for μ , we find the $(p+q)^2$ determinant

$$\begin{vmatrix} -\lambda \gamma_{\alpha\beta} & \gamma_{\alpha b} \\ \gamma_{a\beta} & -\lambda \gamma_{ab} \end{vmatrix} = 0 \quad \begin{matrix} \alpha, \beta = 1, 2, \dots, p, \\ a, b = 1, 2, \dots, q. \end{matrix} \quad (43.62)$$

Multiplying the first p rows by $-\lambda$ and dividing the last q columns by $-\lambda$ we find

$$(-\lambda)^{q-p} \begin{vmatrix} \lambda^2 \gamma_{\alpha\beta} & \gamma_{\alpha b} \\ \gamma_{a\beta} & \gamma_{ab} \end{vmatrix} = 0. \quad (43.63)$$

If we insert another $(p+q)^2$ determinant on the left of (43.63), it will still equal zero. We insert

$$\begin{vmatrix} \mathbf{I}_p & -\gamma_{ab} \gamma_{ab}^{-1} \\ \mathbf{0} & \gamma_{ab}^{-1} \end{vmatrix}.$$

Since the determinant of a product is the product of the determinants, (43.63) becomes

$$(-\lambda)^{q-p} \begin{vmatrix} \lambda^2 \gamma_{\alpha\beta} - \gamma_{\alpha b} \gamma_{ab}^{-1} \gamma_{a\beta} & \mathbf{0} \\ \gamma_{ab}^{-1} \gamma_{a\beta} & \mathbf{I}_q \end{vmatrix} = 0$$

or

$$(-\lambda)^{q-p} \begin{vmatrix} \lambda^2 \gamma_{\alpha\beta} - \gamma_{\alpha b} \gamma_{ab}^{-1} \gamma_{a\beta} \end{vmatrix} = 0. \quad (43.64)$$

This is of order $p+q$ in λ . $q-p$ roots are zero. The others arise in pairs from the $p \times p$ determinant in (43.64). We may write the non-vanishing roots as $\pm\rho_1, \pm\rho_2, \dots, \pm\rho_p$. We choose as the roots those which are not negative and proceed to prove that they are the canonical correlations as we have defined them.

43.21 We have arrived at variates ξ obeying condition (a) at the beginning of 43.20. A simple root of (43.64) substituted in (43.60) gives us the coefficients l and m , except for a multiplicative -1 . For a root of multiplicity t they are determinate except for $t-1$ assignable constants, a result which we take without proof from the theory of algebraic forms.

To complete, we need to prove that the ξ 's in each group are uncorrelated and that, apart from the canonical correlations, any ξ in one group is uncorrelated with any ξ from the other. Suppose we have a root ρ_i and determine the corresponding constants l_i and m_i and hence the pair of corresponding variables ξ_i and η_i . Then we have from (43.60)

$$\sum l_{i\alpha} \gamma_{\alpha\alpha} = \rho_i \sum m_{i\beta} \gamma_{\alpha\beta} \quad (43.65)$$

$$\sum m_{i\alpha} \gamma_{\alpha\alpha} = \rho_i \sum l_{i\beta} \gamma_{\alpha\beta} \quad (43.66)$$

Similar equations obtain for a second pair, say ξ_j and η_j . Between these four variables there are six correlations, of which two are ρ_i and ρ_j . It will be enough to show that the other four vanish. They are

$$E(\xi_i \xi_j) = \sum l_{i\alpha} l_{j\beta} \gamma_{\alpha\beta}, \quad E(\eta_i \eta_j) = \sum m_{i\alpha} m_{j\beta} \gamma_{\alpha\beta}, \quad (43.67)$$

$$E(\xi_i \eta_j) = \sum l_{i\alpha} m_{j\beta} \gamma_{\alpha\beta}, \quad E(\xi_j \eta_i) = \sum l_{j\alpha} m_{i\beta} \gamma_{\alpha\beta}. \quad (43.68)$$

Multiply (43.65) by $m_{j\alpha}$ and sum. In virtue of (43.68) we have

$$E(\xi_i \eta_j) = \rho_i E(\eta_i \eta_j). \quad (43.69)$$

Likewise from (43.66) multiplied by $l_{j\alpha}$ we find

$$E(\xi_j \eta_i) = \rho_i E(\xi_i \xi_j). \quad (43.70)$$

Interchanging i and j , we find from (43.69) and (43.70)

$$\rho_i E(\eta_i \eta_j) = \rho_j E(\xi_i \xi_j), \quad (43.71)$$

and interchanging i and j in this,

$$\rho_j E(\eta_i \eta_j) = \rho_i E(\xi_i \xi_j). \quad (43.72)$$

It follows that unless $\rho_i^2 = \rho_j^2$

$$E(\eta_i \eta_j) = E(\xi_j \xi_i) = 0, \quad (43.73)$$

and in a similar way the other covariances may be shown to vanish.

We have only to round off the proof by showing that if ρ is a root of multiplicity t the property still holds. This follows from the consideration that we may then choose our l 's and m 's to obey certain orthogonal conditions ensuring that

$$E(\xi_i \xi_j) + E(\eta_i \eta_j) = 0.$$

It will then follow from (43.72) that each expectation vanishes unless $\rho_i = \rho_j = 0$; and even in this case (43.69) and (43.70) show that two expectations vanish, and we may then choose our assignable constants so that the others vanish.

43.22 When the variables are put into canonical form the dispersion matrix reduces to

$$\begin{pmatrix} \mathbf{I}_p & \begin{matrix} \rho_1 & & 0 \\ & \rho_2 & \\ & & \ddots \\ 0 & & & \rho_p \end{matrix} & \mathbf{0} \\ \begin{matrix} \rho_1 & & 0 \\ \rho_2 & & \\ & \ddots & \\ 0 & & \rho_p \end{matrix} & \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{q-p} \end{pmatrix} \quad (43.74)$$

with a determinant equal to

$$(1 - \rho_1^2)(1 - \rho_2^2) \dots (1 - \rho_p^2). \quad (43.75)$$

Example 43.5 (from Hotelling (1936), dealing with data of T. L. Kelley)

140 seventh-grade schoolchildren were given four tests in (a) reading speed, (b) reading power, (c) arithmetic speed, and (d) arithmetic power. It is required to find canonical variates for the two reading tests and the two arithmetic tests.

The correlations between the variates were:

	x_1	x_2	x_3	x_4
x_1	1.0000	0.6328	0.2412	0.0586
x_2	0.6328	1.0000	-0.0553	0.0655
x_3	0.2412	-0.0553	1.0000	0.4248
x_4	0.0586	0.0655	0.4248	1.0000

The determinant (43.63) becomes the symmetric determinant

$$\begin{vmatrix} -\lambda & -0.6328\lambda & 0.2412 & 0.0586 \\ & -\lambda & -0.0553 & 0.0655 \\ & & -\lambda & -0.4248\lambda \\ & & & -\lambda \end{vmatrix} \quad (43.76)$$

or
giving
with

$$0.491,370\lambda^4 - 0.078,803,4\lambda^2 + 0.000,362,490 = 0,$$

$$\lambda^2 = 0.155,635 \quad \text{or} \quad 0.004,740$$

$$\lambda = 0.3945 \quad \text{or} \quad 0.0688.$$

To find the transformed variates themselves we use (43.60). For instance, with the root 0.3945 for λ , we have

$$l_1 + 0.6328l_2 + 0.6114m_1 - 0.1485m_2 = 0 \quad (43.77)$$

$$0.6328l_1 + l_2 + 0.1402m_1 - 0.1660m_2 = 0 \quad (43.78)$$

$$-0.6114l_1 + 0.1402l_2 + m_1 + 0.4248m_2 = 0 \quad (43.79)$$

$$-0.1485l_1 - 0.1660l_2 + 0.4248m_1 + m_2 = 0. \quad (43.80)$$

The last equation is linearly dependent on the other three and so adds nothing. In the other three we solve for the ratios of l 's and m 's, finding

$$l_1 : l_2 : m_1 : m_2 = -2.7772 : 2.2655 : -2.4404 : 1.$$

Thus the transformed variates are

$$k_1 \xi_1 = -2.7772x_1 + 2.2655x_2 \quad (43.81)$$

$$k_2 \xi_2 = -2.4404x_3 + x_4, \quad (43.82)$$

where k_1 and k_2 may be chosen so that the variances of ξ_1 and η_1 are unity, if desired. Similar equations with the root 0.0688 will give us a further pair of canonical coordinates. Those we have worked out have the maximum correlation, the other pair having the minimum and therefore being of less interest.

43.23 Standard errors may be obtained in the manner of 43.16. Starting from

$$\sum l_\alpha l_\beta c_{\alpha\beta} = 1 \quad (43.83)$$

$$\sum m_a m_b c_{ab} = 1 \quad (43.84)$$

$$\sum l_\alpha m_a c_{\alpha a} = r, \quad (43.85)$$

we differentiate to find

$$2 \sum c_{\alpha\beta} l_\alpha dl_\beta + \sum l_\alpha l_\beta dc_{\alpha\beta} = 0 \quad (43.86)$$

$$2 \sum c_{ab} m_a dm_b + \sum m_a m_b dc_{ab} = 0 \quad (43.87)$$

$$dr = \sum l_\alpha m_a dc_{\alpha a} + \sum l_\alpha c_{\alpha a} dm_a + \sum m_a c_{\alpha a} dl_\alpha. \quad (43.88)$$

Without loss of generality we may now suppose the variables put into canonical form. All l 's and m 's except l_1 and m_1 vanish and we have

$$2dl_1 + dc_{11} = 0 \quad (43.89)$$

$$2dm_1 + dc_{p+1, p+1} = 0 \quad (43.90)$$

$$dr_1 = dc_{1, p+1} - \frac{1}{2}r_1(dc_{11} + dc_{p+1, p+1}). \quad (43.91)$$

Substituting from the first two in the third of these equations, we find

$$dr_1 = dc_{1, p+1} - \frac{1}{2}r_1(dc_{11} + dc_{p+1, p+1}). \quad (43.92)$$

Similar equations apply to any other simple root, for example

$$dr_2 = dc_{2, p+2} - \frac{1}{2}r_2(dc_{22} + dc_{p+2, p+2}). \quad (43.93)$$

Multiplying (43.92) and (43.93), taking expectations and using (41.98) we find

$$\text{cov}(r_1, r_2) = 0. \quad (43.94)$$

Likewise

$$\text{var } r_1 = \frac{1}{n}(1 - \rho_1^2)^2, \quad (43.95)$$

with similar formulae for the other correlations. It is noteworthy that this is the same as the large-sample formula for an ordinary product-moment correlation.

Hotelling (1936), to whom this derivation is due, showed that if $p = 2$, $q > 2$ and a zero root accordingly has multiplicity t , then nr^2 is distributed as χ^2 with $t-1$ d.f. If a canonical correlation vanishes and $p = q$, (43.95) holds, with the qualification that sample values near the zero root must be allowed to have positive or negative values, or alternatively that the distribution of r is that of the *absolute* value of a normal variate.

Lawley (1959) derives expressions for the third and fourth cumulants of r . He also considers the variance-stabilizing transformation involving $\arctanh r$, but the results are not so satisfactory as for the product-moment correlation in 16.33 (Vol. 1).

43.24 It follows from (43.64) that the ρ^2 are the roots of the determinantal equation

$$|\rho^2 \mathbf{I} - \gamma_{\alpha\beta}^{-1} \gamma_{\alpha b} \gamma_{ab}^{-1} \gamma_{a\beta}| = 0 \quad (43.96)$$

or

$$|\rho^2 \mathbf{I} - \gamma^{-1} \gamma_{12} \gamma_{22}^{-1} \gamma_{21}| = 0, \quad (43.97)$$

where γ_{11} is the matrix of the p variables x_1 to x_p , γ_{22} that of x_{p+1}, \dots, x_{p+q} , γ_{12} is the covariance matrix between the p variables and the q variables, and similarly for γ_{21} . Thus the ρ^2 are the latent roots of the matrix product in (43.97).

43.25 The results of canonical correlation analysis are even more difficult to interpret than those of component analysis. It is best regarded as an exploratory tool which will give us some idea of the structure of the multivariate complex under study, and in any case tells us what can be the maximum amount of correlation between linear functions of the two groups of variables. The literature of the subject has few examples of useful practical application; cf. Barnett and Lewis (1963) for one in educational research. For this reason we will pass rather quickly without proof over some remaining theoretical points.

- For simplicity of exposition we have supposed $q \geq p$. If $q < p$ we simply reverse the roles of the two groups.
- If we insert ML sample values for the matrices in (43.97) we obtain ML estimates of the canonical correlations.
- Looking at the matrices entering into (43.64), we see that one is the dispersion of the p -group and the other (product of three) can be regarded as the contribution from regression of the p -group on the fixed q -group. Thus the theory of regression (42.15–20) applies here. The distribution of the latent roots ρ^2 in (43.97) is that of the λ 's in 41.22–3, provided that the p -group and the q -group are independent, which unfortunately is the case of least interest.
- Bartlett (1947) proposed a test, analogous to that of 43.14, based on the expression of the correlation determinant as the product of p factors $1 - \rho_j^2$. If k canonical correlations have been accepted as non-zero, the criterion for testing that the

others are zero is

$$-\{n-1-k-\frac{1}{2}(p+q+1)+\sum_{j=1}^k r_j^{-2}\} \log \prod_{j=k+1}^p (1-r_j^2), \quad (43.98)$$

which is approximately a χ^2 with $(p-k)(q-k)$ degrees of freedom. Lawley (1959) has investigated this test with reasonably satisfactory results.

- (e) In one other case some progress towards practical application can be made, namely when one canonical correlation is not zero but the others are. See Bartlett (1947).
- (f) Dempster (1966) has considered the removal of bias from estimates of the canonical correlations by Quenouille's method (cf. 17.10, Vol. 2).

Factor analysis

43.26 The methods we have so far discussed in this chapter are designed to examine a system to see what sort of structure it may have. Those we now examine tackle the problem, so to speak, from the other end. We begin with some model of the structure. The problem is to see whether it fits the data and, if not, to modify it until it does.

Specifically, we suppose as usual that we have a $p \times n$ matrix of n observations on a $(p \times 1)$ vector \mathbf{x} . We suppose that the observed x 's are, in fact, linear functions of some underlying variables ζ which are known as factors, there being $m < p$ of them. Thus we have

$$x_j = \sum_{k=1}^m l_{jk} \zeta_k + \varepsilon_j. \quad (43.99)$$

The coefficients l are not now the constants of a rotation to new axes. As in component analysis, they are referred to as factor loadings (a term surviving from early psychological usage for what are more familiar to the statistician as "weights"). The ζ 's are assumed to be independent normal variables with zero mean and unit variance. Since our x -complex, in general, is not representable in fewer than p dimensions, an exact representation of x 's in terms of ζ 's requires an error term ε . As part of the model we suppose that ε_j is independent of ε_k and of all the ζ 's. Our problem is to estimate the constants l and the variances σ_j^2 of the ε 's.

This is not a regression model. Our ζ 's are random variables which we do not regard as fixed quantities like regressors. The relationship is structural in the sense of Chapter 29, Vol. 2.

43.27 The first thing to notice about the model is that it is undetermined. We are representing a p -dimensional complex in terms of $m+p$ random variables. In (43.99) there are pm constants l , mn values of ζ , and pn values of ε . Considered as a set of algebraic equations, (43.99) has many solutions. We have already imposed the condition that the ζ 's are $N(0, 1)$ and we shall also require the ε 's to have zero mean. The question is whether, in conjunction with the conditions of independence among and between ζ 's and ε 's, the problem of estimating pm constants l and p constants σ^2 is determinate.

Since the ζ 's and ε 's are normal, the x 's are also normal. We have then

$$\text{cov}(x_j, x_k) = E\left\{\sum_l l_{jl} \zeta_l + \varepsilon_j\right\} \left\{\sum_l l_{kl} \zeta_l + \varepsilon_k\right\}$$

$$= \sum_{t=1}^m l_{jt} l_{kt}, \quad j \neq k, \quad (43.100)$$

$$\text{var } x_j = \sum_{t=1}^m l_{jt}^2 + \sigma_j^2. \quad (43.101)$$

We may summarize these relations in

$$\gamma = \mathbf{l}\mathbf{l}' + \Sigma \quad (43.102)$$

where \mathbf{l} is the $p \times m$ matrix of coefficients l_{jk} and Σ is the $p \times p$ diagonal matrix of σ_j^2 . The number of dispersions on the left in (43.102) is $\frac{1}{2}p(p+1)$. The number of constants on the right is $p(m+1)$. Thus if $m+1 > \frac{1}{2}(p+1)$ there are not enough relations in (43.102) to determine the constants. We shall, in fact, normalize the constants l by requiring that

$$\sum_{t=1}^p l_{tj} l_{tk} / \sigma_t^2 = 0, \quad j \neq k, \quad (43.103)$$

or equivalently that

$$\mathbf{l}' \Sigma^{-1} \mathbf{l} = \mathbf{J}, \quad (43.104)$$

an $m \times m$ diagonal matrix. This imposes a further $\frac{1}{2}m(m-1)$ conditions on the constants under estimate. The equations will be indeterminate if

$$\frac{1}{2}p(p+1) < p(m+1) - \frac{1}{2}m(m-1)$$

which reduces to

$$(p-m)^2 < p+m. \quad (43.105)$$

We shall therefore assume that the contrary is true.

Example 43.6

The inequality of (43.105) reversed is equivalent to

$$\{(p + \frac{1}{2}) - m\}^2 \geq \frac{1}{4}(8p+1). \quad (43.106)$$

For example, with $p = 5$ our model is indeterminate if m is greater than 2. We should not set up a model of a 5-dimensional complex with more than two factors. For $p = 10$ the largest admissible value of m is 5.

43.28 The reason for imposing the orthogonality conditions (43.103) is as follows. Consider a non-singular orthogonal transformation of the ζ 's to new variables η given by

$$\zeta = \mathbf{M}\eta.$$

The variables η will also be $N(0, 1)$ and independent, and in place of (43.102) we should have

$$\begin{aligned} \gamma &= \mathbf{l}\mathbf{M}(\mathbf{l}\mathbf{M})' + \Sigma = \mathbf{l}\mathbf{M}\mathbf{M}'\mathbf{l}' + \Sigma \\ &= \mathbf{l}\mathbf{l}' + \Sigma. \end{aligned}$$

In short, our ζ 's are indeterminate within an orthogonal transformation. Equation (43.104) resolves the indeterminacy in a convenient way, but there are other methods of doing so.

43.29 If, as we henceforth suppose, $(p-m)^2 > p+m$ we have, in equations (43.102) and (43.103), more equations than constants. We cannot therefore solve them as

a matter of algebra but require some reconciliation procedure. Following Lawley (Lawley and Maxwell, 1963) we shall use the method of maximum likelihood.

One peculiar feature of this situation is that, although we are dealing with ML estimation in normal variation, the sample covariances of the x 's are not estimators of the parent covariances. This is because of the constraints of the estimation represented by (43.102) and (43.103). If we could take the observed c 's as estimators of γ 's we should have simply

$$\begin{aligned} \mathbf{c} &= \hat{\mathbf{l}}\hat{\mathbf{l}}' + \hat{\Sigma} \\ \hat{\mathbf{j}} &= \hat{\mathbf{l}}'\hat{\Sigma}^{-1}\hat{\mathbf{l}}. \end{aligned}$$

As we shall see shortly, the second equation is true but the first is only true of the on-diagonal elements.

We start from the logarithm of the likelihood function

$$\log L = \text{constant} - \frac{1}{2}n \log |\gamma| - \frac{1}{2}n \sum \Gamma_{jk} c_{jk}, \quad (43.107)$$

where Γ is inverse to γ . Substitution from (43.102) for γ gives us a function which we maximize for variations in \mathbf{l} and Σ . Where there is no ambiguity we omit circumflex accents for ease of printing.

Differentiation with respect to σ_t^2 gives us, after some algebra,

$$\Gamma_{tt} - \sum_{j,k} \Gamma_{tj} c_{jk} \Gamma_{kt} = 0 \quad (43.108)$$

which is, summarized for all t , equivalent to

$$\text{diag} (\gamma^{-1} - \gamma^{-1} \mathbf{c} \gamma^{-1}) = \mathbf{0}. \quad (43.109)$$

Differentiation with respect to l_{jk} gives, after some reduction,

$$\sum_t l_{tk} \Gamma_{tj} - \sum_{t,u,v} l_{tk} \Gamma_{tu} c_{uv} \Gamma_{vj} = 0$$

which is the element in the j th row and k th column of

$$\mathbf{l}'\gamma^{-1} - \mathbf{l}'\gamma^{-1} \mathbf{c} \gamma^{-1} = \mathbf{0}. \quad (43.110)$$

To be consistent we must have, as well as (43.108) and (43.110), the equations (43.102) and (43.104) applying to ML estimators,

$$\gamma = \mathbf{l}\mathbf{l}' + \Sigma \quad (43.111)$$

$$\mathbf{J} = \mathbf{l}'\Sigma^{-1}\mathbf{l}. \quad (43.112)$$

From (43.110), postmultiplying by γ , we have

$$\mathbf{l}' - \mathbf{l}'\gamma^{-1}\mathbf{c} = \mathbf{0} \quad (43.113)$$

and hence

$$\mathbf{l}'\mathbf{c}^{-1} = \mathbf{l}'\gamma^{-1}. \quad (43.114)$$

Premultiply (43.109) by $\gamma - \mathbf{l}\mathbf{l}'$, which is equal to Σ . We find

$$\text{diag} (\mathbf{I} - \mathbf{c}\gamma^{-1} - \mathbf{l}\mathbf{l}'\gamma^{-1} + \mathbf{l}\mathbf{l}'\gamma^{-1}\mathbf{c}\gamma^{-1}) = \mathbf{0}$$

which, in virtue of (43.113), reduces to

$$\text{diag} (\mathbf{I} - \mathbf{c}\gamma^{-1}) = \mathbf{0}. \quad (43.115)$$

Multiply this on the right by $\gamma - \mathbf{l}\mathbf{l}'$. We find similarly

$$\text{diag} (\gamma - \mathbf{c}) = \mathbf{0}. \quad (43.116)$$

This is equivalent to the equations

$$\hat{\sigma}_j^2 = c_{jj} - \sum_{k=1}^m \hat{l}_{jk}^2. \quad (43.117)$$

Now from (43.112), postmultiplying by $\mathbf{1}'$, we find

$$\begin{aligned} \mathbf{J}\mathbf{1}' &= \mathbf{1}' \mathbf{\Sigma}^{-1} \mathbf{1}\mathbf{1}' = \mathbf{1}' \mathbf{\Sigma}^{-1} (\mathbf{\gamma} - \mathbf{\Sigma}) \\ &= \mathbf{1}' \mathbf{\Sigma}^{-1} \mathbf{\gamma} - \mathbf{1}'. \end{aligned} \quad (43.118)$$

Thus

$$\mathbf{J}\mathbf{1}' \mathbf{\gamma}^{-1} = \mathbf{1}' \mathbf{\Sigma}^{-1} - \mathbf{1}' \mathbf{\gamma}^{-1}$$

which, in virtue of (43.114), reduces to

$$\mathbf{J}\mathbf{1}' \mathbf{c}^{-1} = \mathbf{1}' \mathbf{\Sigma}^{-1} - \mathbf{1}' \mathbf{c}^{-1}$$

giving

$$\begin{aligned} \hat{\mathbf{J}}\hat{\mathbf{1}}' &= \hat{\mathbf{1}}' (\hat{\mathbf{\Sigma}}^{-1} \mathbf{c} - \hat{\mathbf{\Sigma}} \hat{\mathbf{\Sigma}}^{-1}) \\ &= \hat{\mathbf{1}}' \hat{\mathbf{\Sigma}}^{-1} (\mathbf{c} - \hat{\mathbf{\Sigma}}). \end{aligned} \quad (43.119)$$

43.30 The equations are still troublesome to solve. Recalling that \mathbf{J} is diagonal, we see from (43.119) that its elements are the latent roots of $\mathbf{\Sigma}^{-1}(\mathbf{c} - \mathbf{\Sigma})$. One iterative procedure is to guess some values of $\mathbf{\Sigma}$, determine from (43.119) the latent vectors $\hat{\mathbf{1}}$, substitute in (43.117) to improve the estimate of the σ_j^2 , iterate with these improved estimates in (43.119), and so on. We can estimate $\mathbf{\gamma}$ from (43.111).

The process may, however, converge very slowly (cf. Howe, 1955) and it appears that on occasion the estimates of some of the σ^2 tend to zero. It cannot be said that this subject has been mastered.

43.31 When satisfactory estimates have been obtained, the usual type of likelihood ratio can be used to test whether the number m of factors which have been chosen is satisfactory. Under the hypothesis that there are in fact m factors, the log likelihood is proportional to

$$-\frac{1}{2}n \log |\hat{\gamma}| - \frac{1}{2}n \text{tr}(\mathbf{c}\mathbf{\Gamma}) \quad (43.120)$$

On the hypothesis that the x 's are normally and independently distributed with no errors ε , the sample dispersions are estimators of the parent values and the log likelihood is proportional to

$$\begin{aligned} &-\frac{1}{2}n \log |c| - \frac{1}{2}n \text{tr}(\mathbf{c}\mathbf{c}^{-1}) \\ &= -\frac{1}{2}n \log |c| - \frac{1}{2}np. \end{aligned}$$

Thus the ratio

$$-n \left\{ \log \frac{|c|}{|\hat{\gamma}|} - \text{tr}(\mathbf{c}\mathbf{\Gamma}) + p \right\} \quad (43.121)$$

is distributed approximately as χ^2 . The number of degrees of freedom is the number of constants fitted in the second case less those in the first, which is

$$\frac{1}{2}\{(p-m)^2 - (p+m)\}, \quad (43.122)$$

as we noted in (43.105).

Bartlett (1951) suggested that a better approximation would be obtained by using, instead of n in (43.121), the multiplier

$$n' = n - \frac{1}{6}(2p+11) - \frac{2}{3}m. \quad (43.123)$$

The argument is that this is known to be correct when $m = 0$ (42.12) and is presumably so when $n-m$ and $p-m$ are substituted for n and p .

The rejection of the hypothesis by this test means that more factors are required. The calculation of the criterion (43.121) is tedious, and its value appears to be sensitive to the accuracy of the estimation of the parameters \mathbf{l} and Σ . Lawley and Maxwell (1963) propose an approximate form

$$n' \sum_{j < k} \frac{(c_{jk} - \hat{\gamma}_{jk})^2}{(\hat{\sigma}_j \hat{\sigma}_k)^2} \quad (43.124)$$

where n' is given by (43.123).

43.32 Before the advent of the electronic computer, psychologists were compelled by arithmetical necessity to adopt various devices for obtaining solutions to problems of factor analysis. Some of these can scarcely be said to be more than measures of desperation; others, though difficult to validate with any degree of theoretical rigour, are still useful, if only as providing first approximations from which the iterative solutions of the exact equations can start. Reference for details and numerical illustrations may be made to Harman (1960), Kendall (1961b) and Lawley and Maxwell (1963).

43.33 In factor analysis, as in component analysis, we emerge with expressions giving the variables x as weighted sums of some unobserved—usually unobservable—variables ζ . The main difficulty, as a rule, is to know what the results mean. Psychologists usually try to identify the ζ 's with some factors which they believe to underlie the structure of the system. Application of the same technique to physical systems very often results in weighted sums of variables to which no clear interpretation can be given. For this, and possibly other reasons, we may return to the model to see if any useful modification can be made in it.

43.34 We recall first of all that, to arrive at a unique solution, we imposed an orthogonality condition (43.103) on the factor weights. There is nothing in the model to require this, and, having found the l 's, we are at liberty to transform the ζ 's how we like in the m -dimensional space of ζ 's. We can, in short, *rotate* the factors. We can even transform them to non-independent factors. We have, so to speak, estimated the factor space but are not committed to any particular co-ordinate system within it. There are infinitely many choices, and which we take depends on non-statistical considerations in any particular case. Two criteria suggest themselves:

- (a) To rotate so that as many l 's as possible vanish (or have some minimal property). This is tantamount to invoking some law of parsimony in explaining the x 's in terms of ζ 's—as few ζ 's appear in the relationships as possible.
- (b) To rotate so that some factor loadings are maximized. The object here, as a rule, is similar, for in general, increasing the value of some l 's can only be carried out at the expense of others; but it may also lead to the identification of a factor with a variable x .

From a rather different viewpoint, but with much the same objective, we may impose conditions on the l 's from the outset. For example, we may require that x_1 and x_2

involve only the factor ζ_1 , x_3 to x_4 the factors ζ_1 and ζ_2 , and so on. This procedure is equivalent to putting certain factor loadings equal to zero *a priori* and is said to impose a structure on the system.

43.35 It is evident that in such cases estimational problems become even more severe than in the standard case of 43.29, especially if we permit factors which are themselves correlated. We shall not enter into a discussion of these topics, which indeed have scarcely reached a stage of development in which a critical review of theoretical points is possible. Once again the electronic computer has come to the aid of psychologists by enabling them to specify sundry criteria to determine rotations or structural simplification and to solve the resulting equations, but even the computer may find it hard to provide accurate information about the sampling distributions of the resulting estimators.

43.36 A word of warning may be desirable against attempts at component or factor analysis of matrices which are not obtained by product-moment methods. For instance, the elements of a correlation matrix may be estimated by tetrachoric or biserial coefficients—cf. 26.27–33, Vol. 2. If they are, the matrix is not necessarily positive definite, and in certain cases some of the latent roots may turn out to be negative.

EXERCISES

43.1 A p -variate complex has the following correlation matrix:

$$\begin{pmatrix} 1 & \rho & \rho^2 & . & . & . & \rho^{p-1} \\ \rho & 1 & \rho & . & . & . & \rho^{p-2} \\ . & . & . & . & . & . & . \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & . & . & . & 1 \end{pmatrix}$$

Show that the determinant of the matrix is $(1-\rho^2)^{p-1}$ and hence that the complex cannot be represented in fewer than p dimensions.

43.2 Show that if $\rho > 0$ the complex of Example (not Exercise) 43.1 has one greatest latent root and that all the others are equal. Verify that the sum of the latent roots is p .

43.3 The correlation between variables j and k in a p -variate complex is $1 - |j-k|/p$. Show that the complex cannot be represented in fewer dimensions. For the case $p = 4$ show that the latent roots are $(2 \pm \sqrt{2})/4$ and $(6 \pm \sqrt{26})/4$.

43.4 Show that if the latent roots of a dispersion matrix A are typified by λ_i , those of A^2 are λ_i^2 . Show that for large k the matrix A^k tends to have diagonals which are λ_1^k times the squares of the values of the latent vector ξ_1 , λ_1 being the largest latent root of A .

43.5 In the notation of 43.20, if

$$\begin{aligned} A &= |\gamma_{\alpha\beta}| & B &= |\gamma_{ab}| \\ C &= \begin{vmatrix} 0 & \gamma_{\alpha a} \\ \gamma_{a\alpha} & \gamma_{ab} \end{vmatrix} & D &= \begin{vmatrix} \gamma_{\alpha\beta} & \gamma_{\alpha a} \\ \gamma_{a\alpha} & \gamma_{ab} \end{vmatrix} \end{aligned}$$

show that the *vector correlation coefficient* K defined by

$$K^2 = (-1)^p C / (AB)$$

and the square of the *vector alienation coefficient* Z defined by

$$Z = D / (AB)$$

are invariant under linear transformations of the variables. Show also that

$$K = \pm \prod_{j=1}^p \rho_j;$$

$$Z = \sum_{j=1}^p (1 - \rho_j^2).$$

where the ρ 's are canonical correlations.

(Hotelling, 1936)

43.6 In the notation of the previous exercise, k and z being the sample values of K and Z , show that if the population canonical correlations are all distinct,

$$\text{var } k = \frac{1}{n} K^2 \sum_{j=1}^p \frac{(1 - \rho_j^2)^2}{\rho_j^2}$$

$$\text{var } z = \frac{4}{n} Z^2 \sum_{j=1}^p \rho_j^2$$

$$\text{cov}(k, z) = -\frac{2}{n} KZ \sum_{j=1}^p (1 - \rho_j^2).$$

In particular, when $p = 2$,

$$\text{var } k = \frac{1}{n} \{(1 - K^2)^2 - Z(1 + K^2)\},$$

$$\text{var } Z = \frac{4Z^2}{n} (1 - Z + K^2)$$

$$\text{cov}(k, z) = -\frac{2}{n} KZ (1 + Z - K^2).$$

(Hotelling, 1936)

43.7 In the previous exercise, with $p = q = 2$, show that, in standard measure,

$$k = \frac{r_{13} r_{24} - r_{14} r_{23}}{\{(1 - r_{12}^2)(1 - r_{34}^2)\}^{\frac{1}{2}}}$$

and hence derive a test of the hypothesis that the "tetrad difference" $r_{13} r_{24} - r_{14} r_{23}$ is zero.
(Hotelling, 1936)

43.8 In the notation of Exercise 43.6, show that

$$E(k^\alpha z^\beta) = \prod_{j=1}^p \frac{\Gamma\{\frac{1}{2}(q + \alpha + 1 - j)\} \Gamma\{\frac{1}{2}(n - q + 2\beta - j)\} \Gamma\{\frac{1}{2}(n - j)\}}{\Gamma\{\frac{1}{2}(q + 1 - j)\} \Gamma\{\frac{1}{2}(n - q - j)\} \Gamma\{\frac{1}{2}(n + \alpha + 2\beta - j)\}}.$$

(Girshick, 1939)

43.9 If the latent roots of a multinormal dispersion matrix are all equal, say, to unity, show that a *randomly selected* one of the sample roots has mean unity and variance $(p+1)/n$.
(Girshick, 1939)

43.10 Show that the distribution of the tetrad difference of Exercise 43.7, denoted by u , in samples from uncorrelated parents, is given by

$$dF = \frac{4(n-2) \Gamma^2(\frac{1}{2}n)}{\pi \Gamma^2\{\frac{1}{2}(n-1)\}} \int_u^1 \int_{u/t}^1 \frac{(tv-u)^{n-3} dt dv}{\{(1-t^2)(1-v^2)\}^{\frac{1}{2}}} du.$$

(Girshick, 1939)

43.11 In the notation of 43.29 show that

$$\mathbf{H} = \mathbf{1}' \Sigma^{-1} (\mathbf{c} - \Sigma) \Sigma^{-1} \mathbf{1}$$

is equal to \mathbf{J}^2 and hence is diagonal.

(Lawley and Maxwell, 1963)

43.12 If the "error" variances in a factor analysis are at choice, show that they can be chosen so as to reduce the number of factors required to m if

$$p \geq \frac{1}{2}(p-m)(p-m+1).$$

43.13 Consider a factor analysis with $p = 2$, $m = 1$. Write down the likelihood function and show by differentiation that the ML equations are

$$-c_{12} l_2 + \left(c_{11} - \frac{c_{11} \sigma_1}{l_1} \right) l_2 = 0$$

$$c_{22} \left(1 - \frac{\sigma_2}{l_2} \right) - c_{12} l_2 = 0.$$

Hence that

$$c_{22} l_1^2 = c_{11} l_2^2$$

and thus that

$$\sigma_2 / l_2 = \sigma_1 / l_1.$$

This is an inadmissible result for the free estimation of the four parameters. Explain the reason for its appearance.

43.14 Verify the value for the determinant (43.74) given at (43.75).

CHAPTER 44

DISCRIMINATION AND CLASSIFICATION

44.1 In this chapter we shall be concerned with problems of differentiating between two or more populations on the basis of multivariate measurements. There are three distinct classes of problem which are often confused:

- (a) Discrimination. We are given the existence of two populations and a sample of individuals from each. The problem is to set up a rule, based on measurements from these individuals, which will enable us to allot some new individual to the correct population when we do not know from which of the two it emanates.
- (b) Classification. We are given a sample of individuals, or the whole population, and the problem is to classify them into groups which shall be as distinct as possible. In discrimination the existence of the groups is given; in classification it is a matter to be determined.
- (c) Dissection. We are given a sample or population and wish to divide it into groups, whether the border-lines of subdivision are natural or not.

For example, given a set of individuals from two different races, we may wish to set up a function which will enable us to allocate any freshly observed individual to the correct race. This is a problem of discrimination. Or, given a population of unknown origins, we may wish to see whether they fall into natural classes, natural in this sense meaning that the members in a group are close together in resemblance, but that the members of one group differ considerably from those of another. This is a problem of classification. Finally, given a set of students with observed performances at an examination, we may wish to divide their standard of success into firsts, seconds and thirds, and the points where we effect this division are entirely arbitrary. This is a problem of dissection, and presents itself even where the population is homogeneous.

In this chapter we shall discuss discrimination and classification, but not dissection.

Discrimination

44.2 Before beginning the theory of the subject, it is worth considering whether the problem as we have described it makes practical sense. We are *given* a set of individuals each of which is known with certainty to belong to population *A* or population *B*. If we can acquire this knowledge with certainty for such a group, why not for any new individuals which we may meet? There are at least three types of case providing an answer to this question.

- (a) Lost information. We may require to be able to assign to the correct sex a number of human bones dug up on an archaeological site. While the beings were alive there would have been no problem, but the essential information has crumbled into dust.
- (b) Unattainable information. A sample of hospital records may provide us with data

concerning external symptoms and the existence of internal disease. Our problem is to diagnose the disease from external symptoms without hospitalization; and indeed one of the main objects may be to diagnose and treat at an early stage, so that internal examination is avoided.

(c) Prediction. It may have been found from past experience that we can discriminate between certain types of behaviour, for example of economic systems, on the basis of observations made at a previous point of time. We rely on observations at the present point of time in order to predict the behaviour in the future.

44.3 It is important to note that we shall, in the first case, consider the allocation of an individual to one of two classes without provision for suspended judgement. That is to say, it is mandatory to assign to one of the classes, even at the risk of error. When we make the assignment we may commit two kinds of mistake, according to the population to which we wrongly allocate a given member; and we shall assume in the first instance that the two types are equally important.

Consider then a space W of p dimensions in which a sample member is represented by a point whose co-ordinates are the x -values. The two populations may be imagined as two clusters of points (or continuous densities) which are separate (for otherwise they would be indistinguishable by means of x -values alone) but to some extent overlapping (for otherwise there would be no problem of discrimination). We wish to set up a boundary in the space such that as many as possible of population 1 lie on one side and as many as possible of population 2 on the other. And we require the boundary to have a fairly simple shape. If f_1 and f_2 represent the respective frequency functions, we require our boundary to determine a region R such that

$$\begin{aligned}\int_R f_2 d\mathbf{x} &= \int_{W-R} f_1 d\mathbf{x} \\ &= 1 - \int_R f_1 d\mathbf{x}.\end{aligned}\tag{44.1}$$

This is equivalent to

$$\int_R (f_1 + f_2) d\mathbf{x} = 1.\tag{44.2}$$

This condition means, in effect, that the probabilities of misallocation are the same for the two kinds of error. We further wish to minimize the total error, which is equivalent to minimizing one of the types of error, say

$$\int_R f_2 d\mathbf{x} = \text{minimum}.\tag{44.3}$$

The problem is then to find an unconditional minimum of

$$\int_R \{f_2 - \lambda(f_1 + f_2)\} d\mathbf{x}\tag{44.4}$$

or, equivalently, of

$$\int_R (\beta f_2 - f_1) d\mathbf{x},\tag{44.5}$$

the constants λ or β being determined from (44.2). This is clearly achieved by taking

into R all those points, and only those points, for which $\beta f_2 - f_1 < 0$. The boundary of R is therefore given by

$$f_1/f_2 = \beta,$$

that is to say, by a ratio of likelihoods.

(44.6)

44.4 This is evidently a reasonable criterion. We allocate a member to one population or the other according to its "nearness"; except that "nearness" is not a metrical distance but a nearness in probability. The probability of misclassification for either type of error is given by

$$\int_{f_1/f_2 > \beta} f_2 d\mathbf{x} = \int_{f_2/f_1 > 1/\beta} f_1 d\mathbf{x}.$$

(44.7)

We have, in fact, re-proved the Neyman-Pearson lemma of 22.10, Vol. 2.

44.5 Now suppose that the two populations are multivariate normal with means μ_1 and μ_2 and identical dispersion matrices γ . Apart from constants, the logarithm of the ratio of likelihoods is then, with Γ inverse to γ ,

$$\begin{aligned} -\frac{1}{2} \sum_{j,k} \Gamma_{jk} \{ (x_j - \mu_{1j})(x_k - \mu_{2k}) - (x_j - \mu_{2j})(x_k - \mu_{1k}) \} \\ = -\frac{1}{2} \sum_{j,k} \Gamma_{jk} (\mu_{1j} - \mu_{2j}) x_k - \frac{1}{2} \sum_{j,k} \Gamma_{jk} (\mu_{1j} \mu_{1k} - \mu_{2j} \mu_{2k}). \end{aligned}$$

(44.8)

The second part of this expression is a constant, and without loss of generality we can take our boundary to be determined by

$$\sum_{j,k} \Gamma_{jk} (\mu_{1j} - \mu_{2j}) x_k = \text{constant}.$$

(44.9)

This is a parental form. If we are given a sample with means \bar{x}_1 , \bar{x}_2 and pooled dispersions c_{jk} , the sample boundary function is

$$\sum C_{jk} (\bar{x}_{1j} - \bar{x}_{2j}) x_k = \text{constant}.$$

(44.10)

44.6 The same result may be reached by a different route. Suppose that we determine a linear function

$$X = \sum_{j=1}^p l_j x_j$$

(44.11)

so as to maximize the ratio of between-class to within-class variances, namely

$$\frac{\{ \sum l_j (\mu_{1j} - \mu_{2j}) \}^2}{\sum l_j l_k \gamma_{jk}}.$$

(44.12)

A differentiation with respect to l_j gives us

$$\mu_{1j} - \mu_{2j} = \frac{\left\{ \sum_{j=1}^p l_j (\mu_{1j} - \mu_{2j}) \right\} \sum_k \gamma_{jk} l_k}{2 \sum l_j l_k \gamma_{jk}}$$

from which we have

$$l_k \propto \sum_j \Gamma_{jk} (\mu_{1j} - \mu_{2j}),$$

(44.13)

leading back to (44.9). Since our function X is used only to separate the two populations, not to measure the distance between them, we may multiply it by any convenient constant.

44.7 As we might expect from symmetry, the discriminating hyperplane (44.9) is perpendicular to the line joining the two "centres" whose co-ordinates are μ_1 and μ_2 . To see this, make an orthogonal transformation of the variables to a set which are independent and have unit variance. Since f_1 and f_2 have the same dispersion matrix, the same transformation reduces each to the required form. The discriminating hyperplane becomes

$$\sum (\mu_{1j} - \mu_{2j})x_j = \text{constant},$$

which is perpendicular to the line joining the centres, whose direction cosines are proportional to $\mu_1 - \mu_2$. The distributions f_1 and f_2 are at the same time transformed to spherically symmetric functions, and without loss of generality we may take one co-ordinate axis along the line of means. The integrals giving the errors of classification then reduce to univariate normal integrals and are clearly equal when the discriminating boundary bisects the line of means.

This determines the constant in (44.10). If \bar{X}_1 is the mean of the left-hand side with respect to f_1 , and \bar{X}_2 is that with respect to f_2 , the constant is halfway between them, i.e. is equal to $\frac{1}{2}(\bar{X}_1 + \bar{X}_2)$. Without losing generality, we henceforth assume that $\bar{X}_1 \geq \bar{X}_2$.

Example 44.1 (Fisher, 1936)

Table 44.1 gives the measurements in centimetres of four variables on 50 flowers from each of three varieties of Iris, namely *setosa* (*S*), *versicolor* (*Ve*) and *virginica* (*Vi*). Consider the discrimination of *S* from *Ve*. The variables are:

- x_1 = sepal length
- x_2 = sepal width
- x_3 = petal length
- x_4 = petal width.

The means were (in centimetres):

Variate	<i>Versicolor</i>	<i>Setosa</i>	Difference
x_1	5.936	5.006	0.930
x_2	2.770	3.428	-0.658
x_3	4.260	1.462	2.798
x_4	1.326	0.246	1.080

(44.14)

The pooled sums of squares and products about the means were (in cm^2):

	x_1	x_2	x_3	x_4
x_1	19.1434	9.0356	9.7634	3.2394
x_2		11.8658	4.6232	2.4746
x_3			12.2978	3.8794
x_4				2.4604

(44.15)

THE ADVANCED THEORY OF STATISTICS

318

Table 44.1—Multiple measurements in taxonomic problems

<i>Iris setosa</i>				<i>Iris versicolor</i>				<i>Iris virginica</i>			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	2.5
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	1.9
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	2.1
5.0	3.6	1.4	0.4	6.5	2.8	4.6	1.5	6.5	3.0	5.8	1.8
5.4	3.9	1.7	0.3	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.2
4.6	3.4	1.4	0.2	6.3	3.3	4.7	1.6	4.9	2.5	4.5	2.1
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.7
4.4	2.9	1.4	0.1	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.2	5.2	2.7	3.9	1.4	7.2	3.6	6.1	1.8
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.5
4.8	3.4	1.6	0.1	5.9	3.0	4.2	1.5	6.4	2.7	5.3	2.0
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	1.9
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.1
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.0
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.4
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	2.3
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	1.8
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.2
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	2.3
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	1.5
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.3
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	2.0
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	1.8
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	2.1
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	2.1
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.6
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	1.9
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	7.9	3.8	6.4	2.0
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.4	2.8	5.6	2.2
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.6	6.3	2.8	5.1	1.5
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.5	6.1	2.6	5.6	1.4
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.6	7.7	3.0	6.1	2.3
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.5	6.3	3.4	5.6	2.4
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.4	3.1	5.5	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.0	3.0	4.8	1.8
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.9	3.1	5.4	2.1
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.7	3.1	5.6	2.4
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	6.9	3.1	5.1	2.3
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	5.8	2.7	5.1	1.9
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.8	3.2	5.9	2.3
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.3	5.7	2.5
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.7	3.0	5.2	2.3
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.3	2.5	5.0	1.9
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.3	6.5	3.0	5.2	2.0
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.1	6.2	3.4	5.4	2.3
							1.3	5.9	3.0	5.1	1.8

The inverse matrix is, in cm^{-2} :

	x_1	x_2	x_3	x_4
x_1	0.118,7161	-0.066,8666	-0.081,6158	0.039,6350
x_2		0.145,2736	0.033,4101	-0.110,7529
x_3			0.219,3614	-0.272,0206
x_4				0.894,5506

(44.16)

Questions of degrees of freedom sometimes arise in this class of work, but since the object of the discriminant function is to separate, it can absorb an arbitrary constant in the coefficients. We shall take the total sample number 100 as the number of d.fr. in (44.15). The values in (44.16) are then to be multiplied by 100 to get the inverse of the dispersion matrix.

Using (44.10) we then find for the coefficients

$$l_1 = (118.7161)(0.930) - (66.8666)(-0.658) - (81.6158)(2.798) + (39.6350)(1.080) = -3.115,11.$$

$$l_2 = -18.390,75$$

$$l_3 = 22.210,44$$

$$l_4 = 31.473,74.$$

(44.17)

We may multiply these coefficients by any convenient constant. Taking the coefficient l_1 to be unity would, for example, give us

$$X = x_1 + 5.9037x_2 - 7.1299x_3 - 10.1036x_4. \quad (44.18)$$

The mean value of X for *versicolor*, obtained by substituting means in (44.17), is 66.917. That for *setosa* is -38.424. The mid-point is 14.247. Thus for any value of X above 14.247 we assign to *versicolor*; in the converse case to *setosa*.

44.8 We may calculate approximately the probability of misclassification. We have

$$\begin{aligned} \text{var } X &= \sum l_j l_k \gamma_{jk} \\ &= \sum l_j \gamma_{jk} \Gamma_{km} (\mu_{1m} - \mu_{2m}) \\ &= \sum l_j (\mu_{1j} - \mu_{2j}) \end{aligned}$$

which is estimated using (44.11) by

$$\widehat{\text{var}} X = \bar{X}_1 - \bar{X}_2. \quad (44.19)$$

This is the estimated variance of a single value of X . The variance of half the difference of two mean values, each based on n observations, is $(\bar{X}_1 - \bar{X}_2)/2n$. If we take our critical value of X to be $\frac{1}{2}(\bar{X}_1 + \bar{X}_2)$ the probability of misclassification either way is the probability of exceeding a normal deviation of $\frac{1}{2}(\bar{X}_1 + \bar{X}_2) - \bar{X}_1 = \frac{1}{2}(\bar{X}_2 - \bar{X}_1)$ about zero with variance $(\bar{X}_1 - \bar{X}_2)/2n$.

Example 44.2

In the data of Example 44.1,

$$\bar{X}_1 = 66.917, \bar{X}_2 = -38.424.$$

Hence the error of misclassification is the probability of a deviation of 52.67 or more from the mean of a normal distribution with variance $105.341/2n = 1.053,41$. This is equal to the probability of a deviation of 52.67 with standard deviation 1.023 and is negligible.

44.9 An interesting special case occurs when all the correlations between the x 's are equal. It is not uncommon in biological work for the correlations to be more or less the same in magnitude. If so, we can discriminate on two factors, size and shape, as follows.

As in Example 43.1, it may be shown that if the correlations are all equal to ρ the latent roots of the correlation matrix are given by

$$\lambda_1 = 1 + (p-1)\rho \quad (44.20)$$

$$\lambda_2 = \dots = \lambda_p = 1 - \rho. \quad (44.21)$$

The variation therefore contains one major component, the rest being isotropic. The component corresponding to λ_1 is

$$\zeta_1 = \frac{1}{\sqrt{p}} \sum_{j=1}^p x_j. \quad (44.22)$$

We take a "size" component proportional to this and write

$$Q = \sum x_j = \sqrt{p} \zeta_1 \quad (44.23)$$

so that

$$\text{var } Q = p\lambda_1 = p\{1 + (p-1)\rho\}. \quad (44.24)$$

Among the remaining components no one stands out in advance of the others. Let us then take a set of weights w_j with non-zero mean and define a "shape" component by

$$P = \sum_{j=1}^p \frac{w_j - \bar{w}}{\bar{w}} x_j. \quad (44.25)$$

We find that

$$\text{var } P = \sum \left(\frac{w_j - \bar{w}}{\bar{w}} \right)^2 (1 - \rho). \quad (44.26)$$

Further,

$$\begin{aligned} \text{cov}(Q, P) &= \text{cov} \left(\sum \frac{w_j - \bar{w}}{\bar{w}} x_j, \sum x_j \right) \\ &= \sum_j \frac{w_j - \bar{w}}{\bar{w}} \text{var } x_j + \sum_{j \neq k} \sum \frac{w_j - \bar{w}}{\bar{w}} \text{cov}(x_j, x_k) \\ &= \{1 + (p-1)\rho\} \sum_j \frac{w_j - \bar{w}}{\bar{w}} \\ &= 0. \end{aligned} \quad (44.27)$$

The size and shape components are then uncorrelated.

44.10 To arrive at a discriminator we take

$$w_j = \bar{x}_{1j} - \bar{x}_{2j} \quad (44.28)$$

and look for a discriminator of form

$$X = \alpha Q + P, \quad (44.29)$$

so as to maximize

$$(X_1 - X_2)^2 / \text{var } X.$$

Writing $D_P = P_1 - P_2$, $D_Q = Q_1 - Q_2$ we have then to maximize

$$\frac{(\alpha D_Q + D_P)^2}{\alpha^2 \text{var } Q + 2\alpha \text{cov}(Q, P) + \text{var } P}. \quad (44.30)$$

Using (44.27), we find easily the solution

$$\alpha = \frac{D_Q \text{var } P}{D_P \text{var } Q}. \quad (44.31)$$

Substituting now the weights from (44.28) in the expressions for D_P and D_Q we have

$$D_P = p \sum \frac{(\bar{x}_{1j} - \bar{x}_{2j})^2}{\bar{x}_1 - \bar{x}_2} - (\bar{x}_1 - \bar{x}_2) \quad (44.32)$$

$$\text{var } P = (1 - \rho) \sum \frac{(\bar{x}_{1j} - \bar{x}_{2j})^2}{(\bar{x}_1 - \bar{x}_2)^2} - (\bar{x}_1 - \bar{x}_2) \quad (44.33)$$

$$D_Q = p(\bar{x}_1 - \bar{x}_2) \quad (44.34)$$

$$\text{var } Q = p\{1 + (p-1)\rho\}. \quad (44.35)$$

Substitution in (44.31) then gives α and we find

$$X = \frac{1 - \rho}{1 + (p-1)\rho} Q + P. \quad (44.36)$$

Example 44.3

Consider again the Iris data of Examples 44.1 and 44.2. The correlation matrix is:

	x_1	x_2	x_3	x_4
x_1	1.	.599,513	.636,323	.472,011
x_2		1.	.382,719	.457,988
x_3			1.	.705,258
x_4				1.

(44.37)

The correlations are near enough to equality to justify the use of the foregoing as an approximation. We reduce the variables to zero means and unit variances to give

	Ve	S	$Ve - S$
x_1	1.0628	-1.0628	2.1256
x_2	-0.9551	0.9551	-1.9102
x_3	3.9894	-3.9894	7.9788
x_4	3.4426	-3.4426	6.8852
Sum = Q	7.5397	-7.5397	15.0794 = D_Q

(44.38)

The variance of Q is calculated as the sum of the 16 elements in (44.37) and is given by

$$\text{var } Q = 10.5076. \quad (44.39)$$

The weightings for shape are found from (44.38) by dividing the last column by $\frac{1}{4}(15.0794)$ and subtracting unity, namely are

$$-0.4362, -1.5067, 1.1165, 0.8264. \quad (44.40)$$

For the estimate of P we then find

$$P = (-0.4362 \times 1.0628) + \text{etc.} \\ = 8.2747 \quad (44.41)$$

and again by using (44.38)

$$\text{var } P = 3.0912. \quad (44.42)$$

The covariance of Q and P does not vanish, and we calculate it from (44.38) as 0.36162. Substitution in (44.31) then gives $\alpha = 0.2412$ and our discriminator is

$$X = 0.2412Q + P, \quad (44.43)$$

where, it must be remembered, Q is the sum of the x 's in standard measure and P is the sum weighted by the numbers at (44.40).

Quadratic discriminators

44.11 The linear discriminator (44.10) depends on the assumption that the two populations under comparison have the same dispersions. If this is not so our log likelihood becomes, in an obvious notation, and ignoring constants,

$$\sum \Gamma_{1jk} (x_{1j} - \mu_{1j})(x_{1k} - \mu_{1k}) - \sum \Gamma_{2jk} (x_{2j} - \mu_{2j})(x_{2k} - \mu_{2k}). \quad (44.44)$$

The quadratic terms in x no longer cancel, and our boundary becomes a quadric in p dimensions. This is, in general, an awkward construct to handle, which probably accounts for the fact that quadratic discriminators have not come into general use.

We can make some progress if we reduce the situation to one of size and shape. We now have $p = 2$ and the covariance terms vanish. Expression (44.44) then reduces to a form of type

$$(y - \nu_1)^2 + (x - \nu_2)^2 \quad (44.45)$$

where y and x are linear functions of P and Q and the variances are calculable. The discriminating boundary then becomes an ellipse (in two dimensions) and the situation is tractable. Reference may be made to C. A. B. Smith (1947) for an example.

Testing of a discriminant function

44.12 The process of testing a discriminator needs a little clarification. We may suspect that there is a real difference between the populations but that they are so close together that a discriminator is not very effective; this is measured by the errors of misclassification which, though minimal, may still be large. Or we may think that there is a larger difference between the populations, but our sample size is not large enough to produce a very reliable discriminator; this is really a matter of setting confidence intervals to the function or its coefficients. Or we may fear that the parents are identical and that a discriminant function is illusory.

Tests of discriminant functions have usually been discussed in terms of the last of these possibilities. They are not so much tests of the functions as tests of homogeneity by the use of the function. If heterogeneity is found, the function, *ipso facto*, is significant in the sense that it discriminates between real differences in an optimal

way (except that we use estimators of dispersions and means instead of the unknown parent values). But that way may not be very good even if it is the best available.

44.13 Suppose our two populations have, in fact, identical means. The difference of the means in the discriminator is then U , say, where

$$U = \sum C_{jk} (\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1k} - \bar{x}_{2k}). \quad (44.46)$$

The term $\bar{x}_{1j} - \bar{x}_{2k}$ is the difference of two means, each normally distributed, and is therefore distributed like a mean about zero with twice the variance of a single mean if the sample sizes are the same. It follows—cf. **41.17**—that U is distributed as Hotelling's $T^2/(2n-1)$, based on $2n$ observations. This is equivalent to the distribution of the multiple correlation R^2 when $\mathbf{R}^2 = 0$, by (41.84), and a test can be carried out by an analysis of variance. It seems preferable, however, to test homogeneity in the manner of Chapter 42, which enables us to consider differences in means and dispersions separately.

44.14 Since we observed in **27.28-9** that the null distribution of R^2 does not require the multinormality assumption, it is no surprise that discrimination with two populations does not require it either. Exercise 44.10 shows how we may derive the boundary (44.9) from a LS analysis.

44.15 Let us now extend the discussion to cases where the two types of error are not equally important. There are two ways in which our previous results may require modification:

- (a) It may be known that members from population 1 have a different chance from those of population 2 of being chosen. For example, in selecting a batch of individuals at random to see if they have active tuberculosis, we expect to find many times more healthy than unhealthy patients.
- (b) The consequences of misallocation may be seriously different. It is less dangerous to diagnose a healthy person as unhealthy (because the mistake is likely to be discovered before serious harm is done) than an unhealthy person as healthy (where the reverse may be true).

Let us suppose that the probabilities of emergence of members from our two populations are π_1 and π_2 ($= 1 - \pi_1$). Let us suppose further that we can attach numerical weights to mistakes, a misclassification costing us c_1 and c_2 units respectively. Instead of now minimizing mistakes in number we minimize cost. Then instead of (44.3) we have to minimize

$$c_2 \int_R \pi_2 f_2 d\mathbf{x} + c_1 \int_{1-R} \pi_1 f_1 d\mathbf{x} = c_1 \pi_1 + \int_R (c_2 \pi_2 f_2 - c_1 \pi_1 f_1) d\mathbf{x}. \quad (44.47)$$

This is minimized when the boundary is determined by

$$\frac{c_2 \pi_2 f_2}{c_1 \pi_1 f_1} > 1. \quad (44.48)$$

Thus if we work with $\log f_2/f_1$ as discriminator the effect of introducing the prior

probabilities π and the loss constants c is merely to add a constant to the discriminating function, or equivalently, to displace its critical value by $\log(c_2\pi_2)/(c_1\pi_1)$.

The case of k populations

44.16 When we proceed from discrimination between two populations to discrimination among a number of populations an essentially new point appears. As before, we shall endeavour to divide up the sample space into mutually exclusive regions, one for each population, and allot an observed member to the population in whose region it falls. But the boundaries of the regions are no longer determined by one single discriminant function. Either we must, to achieve optimal properties, have several functions, or, if we must have a single function, we shall have to sacrifice some discriminatory power.

44.17 It will be enough for expository purposes if we consider three populations—the generalization to k is immediate. We will also generalize to the extent of supposing that the probabilities of occurrence of the three populations whose density functions are f_1, f_2, f_3 are respectively π_1, π_2, π_3 ($\pi_1 + \pi_2 + \pi_3 = 1$). If the corresponding regions are R_1, R_2, R_3 a generalization by C. R. Rao of the Neyman-Pearson lemma states that the errors of misclassification are a minimum if the regions are determined by probability ratios which form a simple extension of (44.15). In fact, R_1 is such that $\pi_1 f_1$ is greater than or equal to both $\pi_2 f_2$ and $\pi_3 f_3$; R_2 is such that $\pi_2 f_2 \geq \pi_3 f_3$ and $\pi_1 f_1$; R_3 is such that $\pi_3 f_3 \geq \pi_1 f_1$ and $\pi_2 f_2$.

44.18 In particular, if the three populations are normal with common dispersion matrix γ_{jk} and means $\mu_{1j}, \mu_{2j}, \mu_{3j}$, it follows as in the manner of 44.5 that R_1 must be such that

$$\sum \Gamma_{jk} (\mu_{1j} - \mu_{2j}) x_k \geq \beta_{j2}, \text{ say,} \quad (44.49)$$

$$\sum \Gamma_{jk} (\mu_{1j} - \mu_{3j}) x_k \geq \beta_{j3}, \text{ say.} \quad (44.50)$$

Similarly for the other regions. In the sample, R_1 will be determined as the domain lying between the two hyperplanes (44.49) and (44.50) and including the mean of population 1; and so on. The surfaces of constant weighted probability ratio for populations 1 and 2 are, in fact, given by

$$\log \frac{\pi_1 f_1}{\pi_2 f_2} = \sum \Gamma_{jk} (\mu_{1j} - \mu_{2j}) x_k - \frac{1}{2} \sum \Gamma_{jk} (\mu_{1j} \mu_{2k} - \mu_{2j} \mu_{1k}) + \log \pi_1 / \pi_2. \quad (44.51)$$

In the particular case where all the π 's are equal we may compare the three functions

$$X_1 = \sum \Gamma_{jk} \mu_{1j} x_k - \frac{1}{2} \sum \Gamma_{jk} \mu_{1j} \mu_{1k} \quad (44.52)$$

$$X_2 = \sum \Gamma_{jk} \mu_{2j} x_k - \frac{1}{2} \sum \Gamma_{jk} \mu_{2j} \mu_{2k} \quad (44.53)$$

$$X_3 = \sum \Gamma_{jk} \mu_{3j} x_k - \frac{1}{2} \sum \Gamma_{jk} \mu_{3j} \mu_{3k}, \quad (44.54)$$

and allot a member to R_1, R_2, R_3 according to which of the X 's is the greatest when the sample values are substituted. For if, say, X_1 is the greatest, it follows from (44.51) that $f_1 > f_2$ and $f_1 > f_3$. As usual we may substitute sample values for the unknown parameters in these equations to get an approximate discriminator.

Example 44.4 (C. R. Rao and Slater, 1949)

A number of persons falling into certain neurotic groups obtained the following mean scores in three tests:

Group	Sample size	Mean Score		
		1	2	3
Anxiety state	114	2.9298	1.1667	0.7281
Hysteria	33	3.0303	1.2424	0.5455
Psychopathy	32	3.8125	1.8438	0.8125
Obsession	17	4.7059	1.5882	1.1176
Personality change	5	1.4000	0.2000	0.0000
Normal	55	0.6000	0.1455	0.2182
	256			

(44.55)

The dispersion matrix within groups (250 d.fr.) was

	1	2	3
1	2.300,851	0.251,578	0.474,169
2		0.607,466	0.035,774
3			0.595,094

(44.56)

Its inverse is

	1	2	3
1	0.543,234	-0.200,195	-0.420,813
2		1.725,807	0.055,767
3			2.012,357

(44.57)

For the purposes of this example we will suppose all the π 's to be equal. The six discriminating functions of type (44.49) are then as follows:

	Coefficients			Constant
	x_1	x_2	x_3	
Normal	0.2050	0.1431	0.1947	-0.0931
Personality change	0.7204	0.0649	-0.5780	-0.5107
Anxiety state	1.0515	1.4676	0.2974	-2.5047
Hysteria	1.1678	1.5679	-0.1081	-2.7139
Psychopathy	1.3599	2.4641	0.1336	-4.9182
Obsession	1.7680	1.8611	0.3573	-5.8375

(44.58)

Here the coefficient of x_1 for the normal state is

$$(0.543,234)(0.6000) - (0.200,195)(0.1455) + (-0.420,813)(0.2182) = 0.2050.$$

Suppose, for example, we had a subject with scores 1, 1, 0. The values of the functions, in the order of (44.58), are 0.2550, 0.2746, 0.0144, 0.0218, -1.0942 , -2.2084 . We assign the member to the second group, personality change. In practice, of course, we should do so very tentatively. The normal group is very close and there are only five members in the personality-change group on which the sample discriminators are based.

44.19 From the geometrical viewpoint, the discriminating functions represent hyper-surfaces in p dimensions. Those of **44.18** are, of course, planes. As we have seen, they are not orthogonal to the lines joining the means of distributions. When we have more than two populations the means will not, in general, be collinear. We might, however, find the line of closest fit to the k means, and use variation in the direction of that line as a discriminator. And in fact, if we have k populations we may seek for a function X given by

$$X = \sum_{j=1}^p l_j x_j$$

such that the ratio of variances between and within classes is maximized. It comes to the same thing to maximize the ratio between classes to total variance, as in **44.6**. If \mathbf{A} represents the dispersion matrix between classes and \mathbf{B} the total, this is equivalent to maximizing

$$\lambda = \frac{\sum A_{jk} l_j l_k}{\sum B_{jk} l_j l_k} \quad (44.59)$$

which leads to

$$\sum_k (A_{jk} - \lambda B_{jk}) l_k = 0. \quad (44.60)$$

Thus the largest latent root of $|\mathbf{A} - \lambda \mathbf{B}| = 0$ provides our discriminator. For details reference may be made to Bartlett (1951), E. J. Williams (1952), and Blackith (1960). It appears to us that, in general, the use of one function for discrimination among several populations may be rather Procrustean unless they are so separate that almost any method will yield reasonable results.

Qualitative data

44.20 Our discussion so far has been in terms of measured variables x . In practice we frequently have to deal with situations where some or all of the variables are qualitative. Let us consider the case where they are all qualitative. Suppose there are p of them and the j th variable is divided into s_{jt} categories. In the given sample there will be, say, n_{1jk} , n_{2jk} members in the k th category of the j th variate. If n_1, n_2 are the total sample members we allot a new member in that sub-class to population 1 or population 2 according as

$$\frac{n_{1jk}}{n_1} > \frac{n_{2jk}}{n_2}. \quad (44.61)$$

In short, only the proportions in the class (j, k) are relevant. All the other class frequencies tell us nothing about membership of that class.

44.21 This seems a crude method of procedure, but it is in line with the criterion

we adopted for measured variables; for equation (44.61) merely says that we allocate a new member to the class for which it has the greater probability of occurrence. If we wish to take the matter further we must provide further information:

- (a) If the categories s_{jk} can be ordered in k , that is to say if they follow a natural sequence $s_{j1}, s_{j2}, \dots, s_{jt}$ (as for example in an ordered categorization), it might be possible to utilize information from cells outside the (j, k) th. So far as we know, this has not been attempted.
- (b) If we are prepared to prescribe misclassification costs a somewhat more sophisticated discriminator can be set up on the criterion that a number is to be allocated so as to minimize the cost of misclassification over the whole table. Cochran and Hopkins (1961) examine the procedure. See also Linhart (1959).

44.22 Perhaps the most troublesome case is the one in which some variables are measured and some qualitative. There appears as yet to be no satisfactory theory to deal with this situation. A rather heuristic approach is to construct a score from the qualitative variables (e.g. by representing a dichotomy by 0, 1, a tritomy by -1, 0, 1, etc. and averaging over variables) and then to use that score as a measured variable in conjunction with the other measured variables. Alternatively, a separate discriminator can be constructed from the measured variables for each cell of the qualitative classification—a tedious procedure and one which is apt to reduce the sample numbers for each discriminator to a very low point of reliability. The subject would repay further study.

44.23 Before proceeding to consider distribution-free methods we deal briefly with a few points not yet discussed: (a) reserved judgement, (b) bias in the estimation of misclassification errors, (c) discarding redundant variables.

Reserved judgement

44.24 In many, perhaps most, problems in discrimination it is wise to allow for reserved judgement on borderline cases, and not to insist on an allocation to one of two classes. This means, in geometrical terms, that we wish to divide the sample space into three regions R_1 , R_2 and D_{12} . If a member falls into R_1 we allocate it to population 1; if it falls into R_2 , to population 2. If it falls into D_{12} we admit that the data are insufficient to make a satisfactory judgement. This region, in general, will contain members of both populations fairly intimately mixed up together, and in practice we should probably seek for some other criterion to disentangle them.

It is not difficult to use the linear discriminator to set up the region D_{12} . We merely have to decide on what misclassification probabilities are tolerable, define R_1 and R_2 in terms of them, and assign D_{12} to the remainder of the sample space.

44.25 With more than two populations the number of regions becomes more numerous. With three, for example, we may define regions of doubt D_{12} , D_{23} , D_{31} in terms of the three discriminants, but these will intersect. Thus we may have a region D_{123} wherein we cannot allocate to any population; a region $D_{12.3}$ where we can reject R_3 but cannot allocate as between R_1 and R_2 ; and so on. No particular difficulty

arises, at least with linear discriminators, which divide the sample space into regions with flat boundaries. Problems of interpretation could arise with quadratic or cubic forms.

Bias in the estimation of misclassification errors

44.26 The simplest way of estimating errors of misclassification is to apply the observed discriminator to each member of the sample on which it is based, and to observe the errors in that sample. If we were certain about the parent normality, equality of dispersions, and the accuracy of estimators of means and dispersions used in constructing the actual discriminator, we could estimate the errors theoretically as in Example 44.1. But if we are uncertain about the extent to which our discriminator is sensitive to these assumptions it is better to ascertain the errors in applying it to the observed sample. In fact, this is a procedure we should probably wish to follow in any case, as a precautionary check. It may, however, involve a small bias.

44.27 There are, in practice, two sources of error in the empirical determination of the misclassification error. First of all, we do not know the parent parameters and our discriminant is based on estimates from the sample. On the average, our empirical estimate of error will be greater than the true value. Secondly, our empirical estimate is derived from data to which it has been fitted. Consequently the empirical estimate will, on the average, be less than it would have been had the discriminator been applied to a new sample; but this itself, as we have seen, would be greater than the true value.

Example 44.5 (Cochran and Hopkins, 1961)

The following simple example will exhibit the effect.

Suppose we have two populations P_1 and P_2 and a single variate which can take values a_1 and a_2 . Let the true probabilities that a member in P_1 has the appropriate values be $\pi_1(a_1)$ and $\pi_1(a_2) = 1 - \pi_1(a_1)$, and similarly for $\pi_2(a_1)$ and $\pi_2(a_2)$. If a sample of n_1 from P_1 bears r_1 values of a_1 , the unbiased estimator of $\pi_1(a_1)$ is r_1/n_1 ; and so forth.

The allocation rule will be to place a further observation a_1 in P_1 if the corresponding $r_1/n_1 > r_2/n_2$, and to place an observation a_2 in P_1 if $1 - r_1/n_1 > 1 - r_2/n_2$.

Now consider the case when the true probabilities are given by

	P_1	P_2
a_1	0.9	0.05
a_2	0.1	0.95

Suppose we are given a random sample of one from each population. (This is very trivial but will suffice to make the point.) If the one in P_1 has value a_1 and that in P_2 has value a_2 , the rule for the future is that every observation with a_1 is to be allocated to P_1 , and every one with a_2 is to be allocated to P_2 . The estimated misclassification probability is zero. The actual (but unobserved) probability is $\frac{1}{2}(0.1 + 0.05) = 0.075$. Likewise if the value from P_1 is a_2 and that from P_2 is a_1 the decision rule is reversed. Again the estimated misclassification probability is zero. Actually it is $\frac{1}{2}(0.9 + 0.95) = 0.925$.

If both members bear the value a_1 we should either reserve judgement or toss up for the allocation. In the latter case the estimated and actual probabilities of misclassification are both 0.5. Similarly if both members exhibit a_2 .

Since we have supposed that the two members we have were chosen at random from their respective populations, we can average these probabilities. The results are:

Occurrence	Prob. of occurrence	Prob. of misclassification	
		Estimated	Actual
$P_1(a_1) P_2(a_2)$	$(.9)(.95) = 0.855$	0.0	0.075
$P_1(a_1) P_2(a_1)$	$(.9)(.05) = 0.045$	0.5	0.500
$P_1(a_2) P_2(a_2)$	$(.1)(.95) = 0.095$	0.5	0.500
$P_1(a_2) P_2(a_1)$	$(.1)(.05) = 0.005$	0.0	0.925
		0.07	0.13875

This, of course, is a very extreme case. In sample sizes likely to be worth discussing in practice the bias is much smaller. Further discussion is given by Cochran and Hopkins (1961). See also John (1961). For a much more comprehensive discussion of error rates see Hills (1966).

If, of course, the initial sample on which we base our discriminator was not chosen at random, no quantitative estimate of bias, in general, is possible.

Redundant variables: standard errors

44.28 It is natural to enquire whether all the variables x which appear in our discriminator are necessary. One expects that discarding variables will weaken the discriminatory power, but the loss may be negligible. Looked at from the geometrical viewpoint, if our constellation of points in a p -dimensional space is satisfactorily divided into two by the discriminating hyperplane, the same may be true if we project on to one of the co-ordinate hyperplanes, in which case the variable orthogonal to that plane is redundant.

There are several ways of approaching this problem. It would save a good deal of trouble if we could discard unrewarding variables at the outset without bringing them into the analysis. This, however, is a hazardous operation—cf. some results of Cochran in Exercises 44.6–9. A more direct approach would be to estimate the misallocation errors by omitting certain variables, but this is apt to be tedious if the number of variables is large. We will consider a third approach by deriving the standard error (in large samples) of the coefficients in the linear discriminant.

The coefficient l_k is given by $l_k = \sum_j C_{jk}(\bar{x}_{1j} - \bar{x}_{2j})$.

$$\text{Hence} \quad dl_k = \sum_j \{(\bar{x}_{1j} - \bar{x}_{2j})dC_{jk} + C_{jk}d(\bar{x}_{1j} - \bar{x}_{2j})\}. \quad (44.62)$$

$$\text{Likewise} \quad dl_m = \sum_r \{(\bar{x}_{1r} - \bar{x}_{2r})dC_{rm} + C_{rm}d(\bar{x}_{1r} - \bar{x}_{2r})\}. \quad (44.63)$$

Hence

$$dl_k dl_m = \sum_{j,r} \{(\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1r} - \bar{x}_{2r})dC_{jk}dC_{rm} + (\bar{x}_{1j} - \bar{x}_{2j})C_{rm}dC_{jk}d(\bar{x}_{1r} - \bar{x}_{2r}) + (\bar{x}_{1r} - \bar{x}_{2r})C_{jk}dC_{rm}d(\bar{x}_{1j} - \bar{x}_{2j}) + C_{jk}C_{rm}d(\bar{x}_{1j} - \bar{x}_{2j})d(\bar{x}_{1r} - \bar{x}_{2r})\}. \quad (44.64)$$

Remembering that means are independent of dispersions for normal variation, we have

$$\text{cov}(l_k, l_m) = \sum_{j, r} [(\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1r} - \bar{x}_{2r}) \text{cov}(C_{jk}, C_{rm}) + C_{jk} C_{rm} \text{cov}\{(\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1r} - \bar{x}_{2r})\}]. \quad (44.65)$$

If the two samples are based on n_1, n_2 observations we easily find

$$\begin{aligned} \text{cov}\{(\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1r} - \bar{x}_{2r})\} &= \text{cov}(\bar{x}_{1j}, \bar{x}_{1r}) + \text{cov}(\bar{x}_{2j}, \bar{x}_{2r}) \\ &= \left(\frac{1}{n_1} + \frac{1}{n_2}\right) c_{jr}. \end{aligned} \quad (44.66)$$

We now require the covariance of C_{jk} and C_{rm} .

Let us write temporarily Γ_{jk} for the co-factor of c_{jk} in $|c|$ so that $C_{jk} = \Gamma_{jk}/|c|$. Then

$$\begin{aligned} dC_{jk} &= -\frac{\Gamma_{jk}}{|c|^2} d|c| + \frac{1}{|c|} d\Gamma_{jk} \\ &= -\frac{\Gamma_{jk}}{|c|^2} \sum_{\alpha, \beta} \Gamma_{\alpha\beta} dc_{\alpha\beta} + \frac{1}{|c|} \sum_{\alpha, \beta} \Gamma_{jk, \alpha\beta} dc_{\alpha\beta}, \end{aligned}$$

where $\Gamma_{jk, \alpha\beta}$ is the co-factor of $c_{\alpha\beta}$ in Γ_{jk} ,

$$= \frac{1}{|c|^2} \sum_{\alpha, \beta} \{-\Gamma_{jk} \Gamma_{\alpha\beta} + |c| \Gamma_{jk, \alpha\beta}\} dc_{\alpha\beta}. \quad (44.67)$$

Now in virtue of Jacobi's theorem on determinants

$$|c| \Gamma_{jk, \alpha\beta} = \Gamma_{jk} \Gamma_{\alpha\beta} - \Gamma_{j\beta} \Gamma_{\alpha k}.$$

Hence (44.67) reduces to

$$\begin{aligned} dC_{jk} &= -\frac{1}{|c|^2} \sum_{\alpha, \beta} \Gamma_{j\beta} \Gamma_{k\alpha} dc_{\alpha\beta} \\ &= -\sum_{\alpha, \beta} C_{j\beta} C_{k\alpha} dc_{\alpha\beta}. \end{aligned} \quad (44.68)$$

We now find, on using (41.98), that

$$\text{cov}(C_{jk}, C_{rm}) = \frac{1}{n_1 + n_2} (C_{jm} C_{kr} + C_{jr} C_{km}). \quad (44.69)$$

Substituting from (44.66) and (44.69) in (44.65), we have

$$\begin{aligned} \text{cov}(l_k, l_m) &= \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sum_{j, r} C_{jk} C_{rm} c_{jr} \\ &\quad + \frac{1}{n_1 + n_2} \sum_{j, r} (\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1r} - \bar{x}_{2r}) (C_{jm} C_{kr} + C_{jr} C_{km}) \\ &= \left(\frac{1}{n_1} + \frac{1}{n_2}\right) C_{mk} + \frac{1}{n_1 + n_2} \sum_{j, r} (\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1r} - \bar{x}_{2r}) C_{jm} C_{kr} \\ &\quad + \frac{1}{n_1 + n_2} C_{km} \sum_{j, r} (\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1r} - \bar{x}_{2r}) C_{jr} \end{aligned} \quad (44.70)$$

$$\begin{aligned} &= \left(\frac{1}{n_1} + \frac{1}{n_2}\right) C_{mk} + \frac{1}{n_1 + n_2} l_m l_k + \frac{1}{n_1 + n_2} C_{km} \sum_r l_r (\bar{x}_{1r} - \bar{x}_{2r}) \\ &= \left(\frac{1}{n_1} + \frac{1}{n_2}\right) C_{mk} + \frac{1}{n_1 + n_2} l_m l_k + \frac{1}{n_1 + n_2} C_{km} (\bar{X}_1 - \bar{X}_2). \end{aligned} \quad (44.71)$$

In particular, with $k = m$, and both replaced by j ,

$$\text{var } l_j = \left(\frac{1}{n_1} + \frac{1}{n_2} \right) C_{jj} + \frac{l_j^2}{n_1 + n_2} + \frac{1}{(n_1 + n_2)} (\bar{X}_1 - \bar{X}_2) C_{jj}. \quad (44.72)$$

Example 44.6

Consider again the Iris data of Example 44.1 and let us test l_1 . We have

$$n_1 = n_2 = 50, \quad l_1 = -3.115, 11,$$

$$\bar{X}_1 - \bar{X}_2 = 105.341, \quad C_{11} = 11.871, 61.$$

We find from (44.72)

$$\text{var } l_1 = 0.4749 + 0.0970 + 12.5057 = 13.0776, \quad \text{s.e.}(l_1) = 3.62.$$

The absolute value is less than the standard error, and we should consider whether x_1 can be discarded without serious loss of discriminating power.

In point of fact, as we shall see later, a good discriminator can be based on x_4 alone.

Cochran and Bliss (1948) considered the case where the effect of some variables is abstracted by a covariance technique, and discrimination applied to the remainder. For the method and a worked example, reference may be made to their paper.

For the use of the D^2 statistic and discrimination generally see C. R. Rao (1952).

Distribution-free methods

44.29 We proceed to discuss the possibility of distribution-free methods of discrimination for k populations. Very little work has been done on this subject, and the following sections 44.29–33 should be regarded as suggestions which need to stand the test of experience.

Let us revert to the representation of members as points in a p -dimensional space

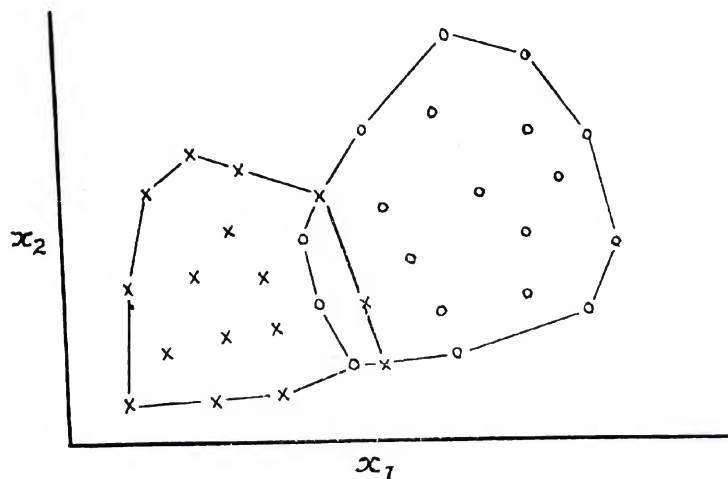


Fig. 44.1 (see text)

whose co-ordinates are the values of the variables x_1, x_2, \dots, x_p . Confining ourselves for the present to two populations, we may think of one population (say A) as represented by crosses and the other (say B) by circles. In two dimensions the picture might look like Fig. 44.1. The crosses have a convex hull which we have drawn in; likewise for the circles. In general these two will have a common domain.

Let us consider the following rule of discrimination:

- (a) If a point falls in the *A*-hull but not in the *B*-hull we assign it to *A*;
- (b) If the point falls in the *B*-hull but not in the *A*-hull we assign it to *B*;
- (c) If the point falls into both hulls we will not assign it to either.

The proposal is plausible but we shall not follow it up for three reasons:

- (i) The determination of the convex hulls is a problem in linear programming which is soluble but takes us outside our present scope;
- (ii) The method gives no guide to the treatment of new points which fall outside both hulls;
- (iii) The method is not truly distribution-free, because non-linear variate transformations do not preserve the planarity of the hull boundaries.

A count of points in the two hulls and their common part is nevertheless useful as giving us a measure of the degree of entanglement of the two populations—a measure, so to speak, of the magnitude of the discrimination problem.

44.30 As a prelude to a distribution-free method, consider again Table 44.1, data for *setosa* and *versicolor*.

The petal width of *setosa* has a mean value of 0.246 and a range of 0.2–0.6 (variance 0.0109). That of *versicolor* has a mean of 1.326 and a range of 1.0 to 1.8 (variance 0.0383). On this showing, as we have already remarked, petal width would be a perfectly good discriminator in itself. If we allot a new member to *setosa* or *versicolor* according as petal width is less than or exceeds, say, 0.9, we shall rarely make a mistake even if the variates are normal.

44.31 The method we propose may be illustrated on the discrimination of *versicolor* against *virginica*. A casual inspection of the data shows what can be confirmed by tabulation, that the two differ more on petal length PL and petal width PW than on sepal length or width. We form a frequency distribution for PL and PW as in Table 44.2.

We observe that on PL the two distributions overlap in the range 4.5–5.1. Outside this range there are 29 cases of *versicolor* and 34 cases of *virginica*. On PW there is overlap in the range 1.4–1.8, 28 cases of *versicolor* and 34 of *virginica* lying outside it. The total of cases lying outside the common range being 63 for PL and 62 for PW, we shall take as our first discriminating variable PL.

We then lay down the following rule of discrimination:

$PL \leq 4.4$ allot to *versicolor*

$PL \geq 5.2$ allot to *virginica*

$4.5 \leq PL \leq 5.1$ refer to next variable.

(44.73)

There are 37 cases for which PL lies in the common range 4.5–5.1. We take these cases out of Table 44.2 and construct a distribution for them in respect of PW, as in Table 44.3.

Table 44.2—Frequency distributions of petal length and petal width for *Iris versicolor* and *Iris virginica*

Variate values	Petal length		Variate values	Petal width	
	Vers.	Virg.		Vers.	Virg.
4.3	25				
4.4	4		1.0	7	
4.5	7	1	1.1	3	
4.6	3	—	1.2	5	
4.7	5	—	1.3	13	
4.8	2	2	1.4	7	1
4.9	2	3	1.5	10	2
5.0	1	3	1.6	3	1
5.1	1	7	1.7	1	1
5.2		2	1.8	1	11
5.3		2	1.9		5
5.4		2	2.0		6
5.5		3	2.1		6
5.6		6	2.2		3
5.7		3	2.3		8
5.8		3	2.4		3
5.9		13	2.5		3
	50	50		50	50

Table 44.3—Frequency distribution of 37 cases not distinguished by PL

Variate values	Petal width	
	Vers.	Virg.
1.2	1	
1.3	2	
1.4	4	
1.5	9	2
1.6	3	—
1.7	1	1
1.8	1	5
1.9		3
2.0		3
2.1		—
2.2		—
2.3		1
2.4		1
	21	16

Proceeding as before, we see that there is a common range for PW of 1.5–1.8. We therefore add to the rule (44.73):

$$4.5 \leq \text{PL} \leq 5.1$$

PW \leq 1.4 allot to *versicolor*

PW \geq 1.9 allot to *virginica*

1.5 \leq PW \leq 1.8 proceed to next variable. (44.74)

This leaves 22 cases undecided. PW has discriminated 63 cases and PL a further 15. We now refer to the 22 undecided cases on sepal length SL and sepal width SW.

Table 44.4—Frequency distributions of 22 cases not distinguished by PL and PW

Variate values	Sepal length		Variate values	Sepal width	
	Vers.	Virg.		Vers.	Virg.
4.9	—	1	2.2	1	1
—	—	—	2.3	—	—
5.4	1	—	2.4	—	—
5.5	—	—	2.5	1	1
5.6	1	—	2.6	—	—
5.7	—	—	2.7	1	1
5.8	—	—	2.8	1	2
5.9	1	1	2.9	1	—
6.0	3	2	3.0	3	3
6.1	—	1	3.1	2	—
6.2	1	1	3.2	2	—
6.3	2	2	3.3	1	—
6.4	1	—	3.4	1	—
6.5	1	—			
6.6	—	—			
6.7	2	—			
6.8	—	—			
6.9	1	—			
	14	8		14	8

Table 44.5—Distribution of 16 cases not distinguished by PL, PW, SW

Variate values	Sepal length	
	Vers.	Virg.
4.9	—	1
—	—	—
5.4	1	—
5.5	—	—
5.6	1	—
5.7	—	—
5.8	—	—
5.9	—	1
6.0	2	2
6.1	—	1
6.2	1	1
6.3	1	2
6.4	—	—
6.5	1	—
6.6	—	—
6.7	1	—
	8	8

For SL there are only 5 cases out of 14 lying outside the common range. For SW there are 6. We therefore take SW as our next discriminator and add to (44.74)

$$4.5 \leq PL \leq 5.1$$

$$1.5 \leq PW \leq 1.8$$

$SW \geq 3.1$ allot to *versicolor*

$SW < 3.1$ proceed to next variable. (44.75)

Our third variable discriminates a further 6, making 84 altogether and leaving 16 undecided. For these 16 the distribution on SL is given in Table 44.5. For what it is worth we may now add to (44.75)

$$4.5 \leq PL \leq 5.1$$

$$1.5 \leq PW \leq 1.8$$

$$SW < 3.1$$

$SL \geq 6.4$ allot to *versicolor*

$SL \leq 5.3$ allot to *virginica*

$5.4 \leq SL \leq 6.3$ undecided. (44.76)

This leaves us with 87 cases decided and 13 undecided. No further discrimination is possible.

44.32 The general method will now be clear. It is completely distribution-free, depending only on the rank order of the variate values. It brings up one by one the variables which are *prima facie* most important in the discrimination. It involves no arithmetic other than counting.

On the other hand, the discrimination which results is not necessarily optimal. Looked at from the geometrical viewpoint, instead of a plane boundary as in Example 44.1, we have a step-wise boundary. The discrimination on the first variable rules off three domains by hyperplanes orthogonal to that variable. The second variable rules off similarly in the region of indecision left by the first; and so on. It is possible that an optimal method based on distributions may leave a smaller residuum of undecided cases than the one we propose; but it can do so, of course, only at the expense of sacrificing the distribution-free nature of the procedure.

Differences in dispersion

44.33 We may add a final word on the problem of discrimination when populations differ in dispersion but not in means. It is easier to point to the problem than to suggest a solution. Consider, for example, Fig. 44.2, where the populations have the same mean but different dispersions. There are clearly areas where discrimination is possible, but the foregoing methods fail to reveal them.

If the configuration was the same but the figure was rotated through 45 degrees, we should arrive at meaningful results by the rank order method. The lines KK' , LL' would rule off domains outside of which crosses were, so to speak, dominant and likewise MM' , NN' would define domains for the circles. The rectangle in the middle would be a zone of indecision, which would inevitably be large owing to the nature of the data.

A heuristic procedure in such cases would be to rotate the axes (for measurable

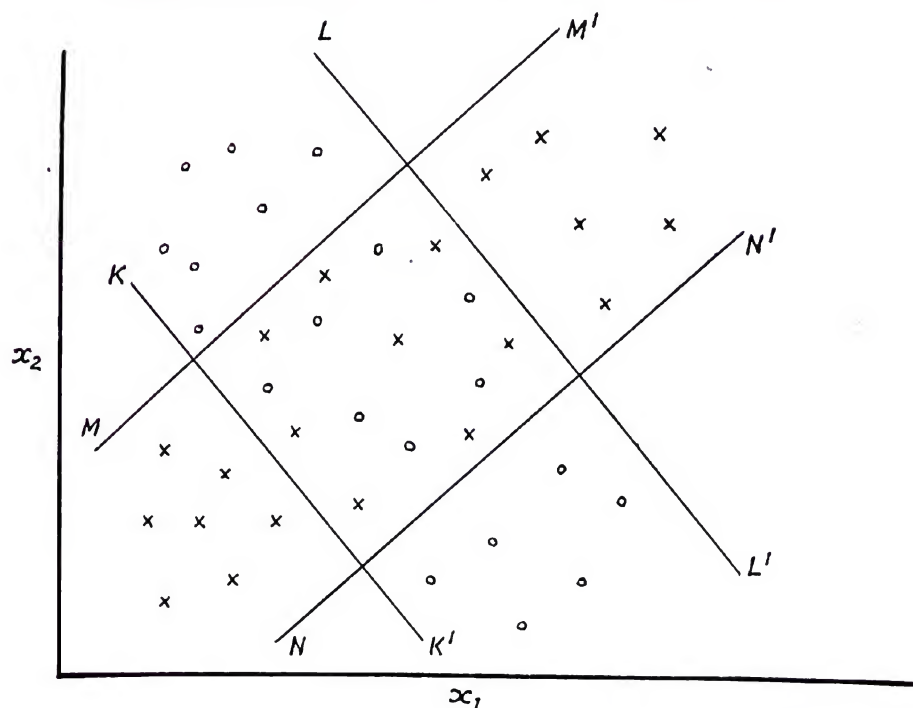


Fig. 44.2 (see text)

variables), say, by a transformation to principal components. Lubischew (1962) has discussed the problem in a biological context.

44.34 One difficulty of the foregoing method stems from its sensitivity to outlying values. As we have explained it, only non-overlapping regions are accepted for discrimination; for variables of effectively infinite range there tends to be more overlap as sample size increases. It might, therefore, be preferable to accept some misclassification from the outset by permitting overlap up to a specified amount; or to fit univariate distributions and estimate the cut-off points to a specified degree of overlap. Much more remains to be done in this field.

Classification

44.35 The problem of classification, as we define the word, is one of determining from empirical evidence whether individuals "group" or "cluster." There are two different ways of looking at this problem, corresponding to the two kinds of space in which we represent the data.

- (a) Given, as usual, a $p \times n$ vector of observations, let us consider the n sample points in the p -dimensional Euclidean space determined by the p variables. If these points, to some acceptable definition, fall into clearly distinguishable groups, we may say that the n individuals may be *classified* into those groups. Their "nearness" is to be considered as a function of the variate values which they bear.
- (b) In the alternative p -space embedded in an n -space the variables are represented by vectors. There is some interest in how far these vectors cluster, as we have seen in canonical analysis. In this case we are concerned with the extent to which the variables cluster, not the individuals.

It would be convenient, though it is not general practice, to refer to the first type as *classification analysis* and the second as *cluster analysis*. In the first we accept the variables and try to classify individuals; in the second, which is perhaps logically anterior to the first, we are interested in the variables, to see, for example, whether they are all necessary and which are the more important for the purpose in hand.

44.36 In either case our primary difficulty is to define what we mean by "group" or "cluster." There are several ways of doing so, but they all rest on the notion of "nearness" or "distance." The consequence is that we have to set up some kind of metric to determine the distance between two points, and then decide on a distance within which two points are "near."

For cluster analysis an obvious distance function of x_j and x_k is the correlation ρ_{jk} . We can regard this either as the cosine of the angle between the vectors or as the cosine of the distance between the end-points of the vectors on the unit hypersphere. The correlation matrix ρ then sets up our distance function. We have only to decide what values constitute nearness and how we use them to define a cluster.

44.37 Suppose we decide that points with $\rho \geq 0.7$ are near together. One manner of procedure is then as follows: scan the correlation matrix for pairs with correlation ≥ 0.7 . If there are none, no cluster exists. In the contrary case take one pair, say x_j, x_k . Examine the correlations of other variables with these two. If there is an x_l such that the average correlation (three values) between x_j, x_k, x_l is ≥ 0.7 add x_l to the cluster. Proceed if possible to find a fourth such that the average ρ (6 values) ≥ 0.7 ; and so on until the process fails. The resulting vectors are a cluster. Putting these on one side, repeat the procedure with the remaining variables; and so on until the set is exhausted.

The procedure is fairly easy to apply for a number of vectors of reasonable size—and in practice the number rarely exceeds 50. But it may not be unique in the sense that where we have a choice of starting pairs the ultimate result may depend on which we choose. If computational facilities were available, it might be possible to split the p vectors into groups in all possible ways, the number of non-unitary partitions of p , and examine the clustering within each partition. But this would probably overtax the capacity of the largest computer.

For some further studies see Tryon (1939) and Fortier and Solomon (1966). The methods of cluster analysis have not been much used by statisticians and are worthy of further study, for example in the discarding of redundant variables in regression analysis, structure analysis, discriminant analysis and, indeed, in multivariate analysis generally. It must be remembered that correlation coefficients are quantities of a highly summary kind, and it is prudent, as a preliminary in all these cases, to draw some of the bivariate scatter diagrams in order to get an overall view of the nature of the variation.

44.38 The method of cluster analysis by correlations has the advantage of being independent of the scale of measurement of any particular x_j . We are, so to speak, concerned with the number of components p , not the way in which any one is measured. Such a method is not distribution-free, but if any worry is felt about the non-normality of the data, the original variables can be replaced by ranks and the correlation procedure

still applies. In fact, we can extend the method to cover qualitative data, provided that the categories in which they occur are orderable (see, for example, 33.36, Vol. 2). The metric, one might claim, is a natural one.

But when we consider the grouping problem of n points in a p -space such considerations no longer apply. "Distances" may be greatly affected by altering the scale of one of the variables and, indeed, can be assigned almost any values we like by stretching scales in the right way. Sometimes the difficulty can be overcome, or at least partially met, by an initial standardization. We prefer, however, a distribution-free method based on ranks.

44.39 We shall now set up a distance function, not between variables but between individuals. Thus, if the variate values of the j th and k th members are $x_{1j}, x_{2j}, \dots, x_{pj}$ and $x_{1k}, x_{2k}, \dots, x_{pk}$, we require a measure of correlation between them. To compute a correlation based on the values as they stand would be nugatory; for example, changing the sign of one vector variable would alter the value of product-moment correlations.

We therefore replace the n values of any component x_j by a set of ranks from 1 to n . These ranks may be tied for any set of members exhibiting the same value of x_j ; and in particular qualitative data in ordered categories may be regarded as tied rankings, so that our method has a very general application. The $p \times n$ matrix of rank values typified by $r_{\alpha\beta}$, $\alpha = 1, 2, \dots, p$; $\beta = 1, 2, \dots, n$ replaces the original matrix.

For each of the pairs of sample members we calculate the function, analogous to a chi-squared measure,

$$D_{jk} = \sum_{\alpha=1}^p \frac{(r_{\alpha j} - r_{\alpha k})^2}{\text{var } r_{\alpha}}. \quad (44.77)$$

The variance of a set of n ranks depends on the number of ties present. If there are ties of t_1, t_2, \dots , etc. members, we have

$$\text{var } r_{\alpha} = \frac{1}{12n} \{(n^3 - n) - \sum(t^3 - t)\}. \quad (44.78)$$

(Cf. Exercise 44.4.)

44.40 A practical difficulty, as in most classification procedures, arises from the number of pairs which can be chosen from n members, namely $\frac{1}{2}n(n-1)$. Thus, for a sample of 100 there are 4950 pairs, each with a value of D_{jk} . To proceed in the manner of 44.37 and form groups by adding one member at a time is a sufficiently complicated exercise to require a computer; but it presents no theoretical problems.

Example 44.7

A heuristic procedure which gives at least a preliminary idea of the extent of the grouping may be illustrated on some of the data of Table 44.1 (Kendall, 1966). A table was constructed from the figures for *versicolor* and *virginica* by regarding them as a single sample of 100 of unknown origin. The first two variables, sepal length and sepal width, were ranked in the ordinary way from 1 to 100. Petal length was split into four categories, values < 4 , ≥ 4 and < 5 , ≥ 5 and < 6 , and ≥ 6 . Petal width

was condensed into two categories < 2 and ≥ 2 —a very heavily tied ranking. The 4950 values of D_{jk} were computed and used to sort the 100 members into classes.

The data gave two well-defined classes, comprising 58 (A) and 25 (B) members, none of them overlapping. Of the 17 remaining there was another group round a further pair, but in fact this group comprised 30 members, 9 new ones from the 17, and 21 in common with A . It was therefore decided to amalgamate the 9 with A to form a group A' of 67 members. B remained with 25. The remaining 8 did not fall into a clearly defined class.

There was thus fairly clear evidence of two classes, and only two. But whereas B contained only *versicolor*, A' contained 48 *virginica* and 19 *versicolor*. On this basis we should correctly arrive at the number of classes, but misclassify 19 and leave 8 doubtful. In the analogous problem of discrimination we decided 87 cases and left 13 doubtful. However, in the present example, we sacrificed a good deal of information by grouping petal length and petal width, so the results are not discordant. Reference may be made to Kendall (1966) for details.

44.41 The subject is far from being exhausted. There are several ways of dividing members into groups, even when a suitable distance metric has been decided upon. For example, it is possible to consider a classification based on the intra-group distance in relation to the between-group distance. Wald (1944) proposed a statistic of this kind which is closely akin to the discriminant function.

As a final comment, we would remark that, under the influence of the papers by Fisher (1936) and Wald (1944), statisticians have tended to approach the problems of discrimination and classification by looking for a single *function* of the variables. This appears to us to be a procedure which, in many circumstances, may be too restrictive. What is required is an allocation *rule* or set of rules; and this may or may not make use of a linear function, or a single function, of the variables.

EXERCISES

44.1 Taking x_1 and x_2 as sepal length and sepal width, respectively, for the data of *Iris setosa* and *versicolor* in Example 44.1, show that the linear discriminant function is

$$x_1 - 1.236 x_2$$

and that this is nearly as good as the four-variable discriminator of the example.

44.2 Show that the discriminating boundary given by (44.9) may be written

$$U \equiv \mathbf{x}' \mathbf{V}^{-1} (\mu_1 - \mu_2) - \frac{1}{2} (\mu_1 + \mu_2)' \mathbf{V}^{-1} (\mu_1 - \mu_2) = 0$$

where \mathbf{V} is the parental dispersion matrix.

Show that if \mathbf{x} is distributed as $N(\mu_1, \mathbf{V})$, U has mean equal to

$$\frac{1}{2} (\mu_1 - \mu_2)' \mathbf{V}^{-1} (\mu_1 - \mu_2) = \frac{1}{2} \alpha, \text{ say,}$$

and variance α .

If \mathbf{x} is distributed according to $N(\mu_2, \mathbf{V})$, show that U has mean $-\frac{1}{2} \alpha$ and variance α .

(T. W. Anderson, 1958. The distribution when sample estimates are inserted for μ and \mathbf{V} is very complicated but asymptotically is the same. See Wald (1944), Sitgreaves (1952), T. W. Anderson (1951), and John (1961).)

44.3 $x, x_{11}, x_{21}, \dots, x_{n_1 1}$ are drawn from population 1 which is $N(\mu_1, V)$. $x_{12}, x_{22}, \dots, x_{n_2 2}$ are drawn from population 2 which is $N(\mu_2, V)$. Consider this against the alternative hypothesis that $x_{11}, \dots, x_{n_1 1}$ are drawn from population 1 and $x, x_{12}, \dots, x_{n_2 2}$ from population 2. Show that the likelihood ratio for testing the composite hypothesis is

$$\frac{1 + \frac{n_2}{n_2 + 1} (\mathbf{x} - \bar{\mathbf{x}}_2)' V^{-1} (\mathbf{x} - \bar{\mathbf{x}}_2)}{1 + \frac{n_1}{n_1 + 1} (\mathbf{x} - \bar{\mathbf{x}}_1)' V^{-1} (\mathbf{x} - \bar{\mathbf{x}}_1)}$$

(T. W. Anderson, 1958)

44.4 In a set of n ranks the ranks $p_{k+1}, p_{k+2}, \dots, p_{k+t}$ are tied and allotted the rank $p_k + \frac{1}{2}(t+1)$. Show that their sum of squares is reduced by $\frac{1}{12}(t^3 - t)$. Deduce the formula for the variance of a ranking with t_1, t_2, \dots, t_m ties:

$$\text{var } r = \frac{1}{12n} \left\{ (n^3 - n) - \sum_{j=1}^m (t_j^3 - t_j) \right\}.$$

44.5 The data for *versicolor* and *virginica* may be classified by petal length and petal width in the following ordered contingency table. (For petal width "small" means < 1.5 ; for petal length "small" means < 4.0 , "medium" means ≥ 4 and < 5 , "large" means ≥ 5 . Figures to the left of the colon refer to *versicolor*, those to the right to *virginica*.)

Petal width

		Small	Large	Totals
Petal length	Small	11:0	0:0	11:0
	Medium	24:0	13:6	37:6
	Large	0:1	2:43	2:44
	Totals	35:1	15:49	50:50

Show that if a new member is assigned on the basis of a majority in the cell frequency in which it falls, the probability of misclassification can be estimated at 8 per cent. Consider the meaning and reliability of this figure.

44.6 A vector \mathbf{x} is p -variate normal and samples are drawn from each of two populations with identical dispersion matrices. Each variable in each population is scaled to have unit variance, and those in the first population have zero mean. The other means are given by $\delta_j, j = 1, 2, \dots, p$, and are taken as positive (by a change of sign of x_j if necessary).

If x_1 alone is used as discriminant show that the probability of misclassification, errors of either kind being equally important, is

$$\frac{1}{\sqrt{(2\pi)}} \int_{\frac{1}{2}\delta_1}^{\infty} \exp(-\frac{1}{2}x^2) dx$$

(Cochran (1964). The suggestion has been made that if this probability is large, say $\delta_j < \frac{1}{2}$, the variable should be discarded as a poor discriminator.)

44.7 In the previous exercise, if all p variables are independent, show that the best combined discriminator, scaled to unit variance, is

$$\sum_{j=1}^p \delta_j x_j / \sqrt{\sum \delta_j^2}.$$

For two independent variates x_1, x_2 with values δ_1, δ_2 ($\delta_1 > \delta_2$) show that an observation on the first variate is equivalent to m observations on the second variate, where $m = \delta_1^2/\delta_2^2$.

44.8 In the previous exercise, let the variables x_1, x_2 have correlation ρ . By considering the independent variables $x_2 - \rho x_1$ and x_1 show that if $\delta_2 = f\delta_1$, $0 \leq f \leq 1$, the correlation improves the discrimination over what it would be if the variables were independent, provided that

$$\frac{(f - \rho)^2}{1 - \rho^2} > f^2.$$

Hence show that a negative correlation always helps the discrimination but that a positive correlation is harmful unless $\rho > 2f/(1 + f^2)$.

(Cochran, 1964)

44.9 Continuing the previous exercise, suppose that the correlations between any pair x_j, x_k are the same and equal to ρ . Show that if ρ is negative, a discriminator based on all p variables is better than it would be if they were independent; but that if ρ is positive this is not so unless

$$\rho > \frac{\left(\sum_{j=1}^p \delta_j\right)^2 - \sum_{j=1}^p \delta_j^2}{(p-1) \sum_{j=1}^p \delta_j^2}.$$

(Cochran, 1964)

44.10 Show that discrimination with two populations may be formally represented as a Least Squares regression analysis in which the dependent variable y can assume only two values, namely $m = n_2/(n_1 + n_2)$ for the n_1 members of the first population and $(m-1)$ for the n_2 members of the second. This yields the boundary in Exercise 44.2 without the assumption of multinormality.

44.11 Discuss the approach of Exercise 44.10 for more than two populations, and show why it breaks down.

CHAPTER 45

TIME-SERIES: GENERAL

45.1 Observations on a phenomenon which is moving through time generate an ordered set known as a time-series. The values assumed by a variable at time t may

Table 45.1—Annual yields per acre of barley in England and Wales from 1884 to 1939
(Data from the Agricultural Statistics)

Year	Yield per acre (cwt)	Year	Yield per acre (cwt)	Year	Yield per acre (cwt)	Year	Yield per acre (cwt)
1884	15.2	1898	16.9	1912	14.2	1926	16.0
85	16.9	99	16.4	13	15.8	27	16.4
86	15.3	1900	14.9	14	15.7	28	17.2
87	14.9	01	14.5	15	14.1	29	17.8
88	15.7	02	16.6	16	14.8	30	14.4
89	15.1	03	15.1	17	14.4	31	15.0
90	16.7	04	14.6	18	15.6	32	16.0
91	16.3	05	16.0	19	13.9	33	16.8
92	16.5	06	16.8	20	14.7	34	16.9
93	13.3	07	16.8	21	14.3	35	16.6
94	16.5	08	15.5	22	14.0	36	16.2
95	15.0	09	17.3	23	14.5	37	14.0
96	15.9	10	15.5	24	15.4	38	18.1
97	15.5	11	15.5	25	15.3	39	17.5

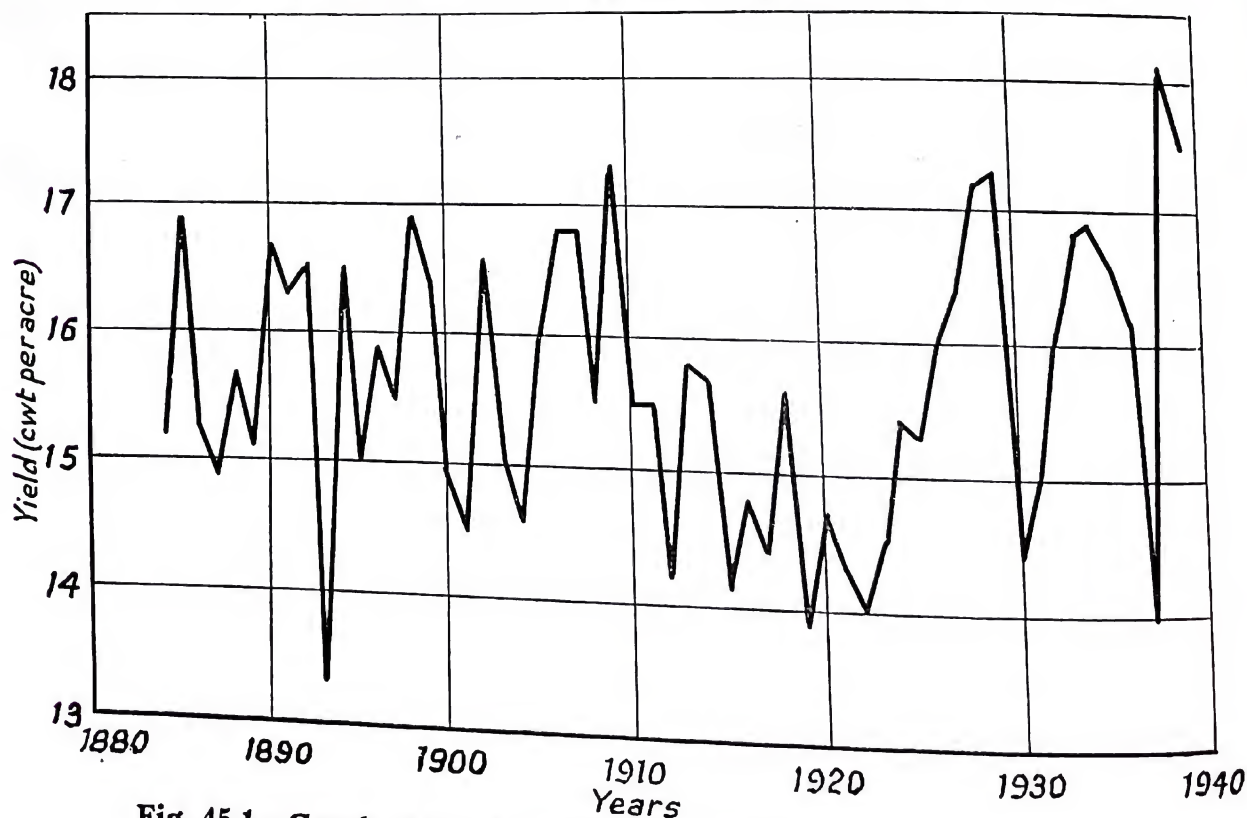


Fig. 45.1—Graph of the data of Table 45.1 (barley yields per acre)

Table 45.2—Total annual rainfall at London in inches, for each year from 1813 to 1912
(Data from D. Brunt, *Phil. Trans.*, A, 225, 247, 1925)

Year	Rainfall (inches)	Year	Rainfall (inches)	Year	Rainfall (inches)	Year	Rainfall (inches)
1813	23.56	1838	21.63	1863	21.59	1888	27.74
14	26.07	39	27.49	64	16.93	89	23.85
15	21.86	40	19.43	65	29.48	90	21.23
16	31.24	41	31.13	66	31.60	91	28.15
17	23.65	42	23.09	67	26.25	92	22.61
18	23.88	43	25.85	68	23.40	93	19.80
19	26.41	44	22.65	69	25.42	94	27.94
20	22.67	45	22.75	70	21.32	95	21.47
21	31.69	46	26.36	71	25.02	96	23.52
22	23.86	47	17.70	72	33.86	97	22.86
23	24.11	48	29.81	73	22.67	98	17.69
24	32.43	49	22.93	74	18.82	99	22.54
25	23.26	50	19.22	75	28.44	1900	23.28
26	22.57	51	20.63	76	26.16	01	22.17
27	23.00	52	35.34	77	28.17	02	20.84
28	27.88	53	25.89	78	34.08	03	38.10
29	25.32	54	18.65	79	33.82	04	20.65
30	25.08	55	23.06	80	30.28	05	22.97
31	27.76	56	22.21	81	27.92	06	24.26
32	19.82	57	22.18	82	27.14	07	23.01
33	24.78	58	18.77	83	24.40	08	23.67
34	20.12	59	28.21	84	20.35	09	26.75
35	24.34	60	32.24	85	26.64	10	25.36
36	27.42	61	22.27	86	27.01	11	24.79
37	19.44	62	27.57	87	19.21	12	27.88

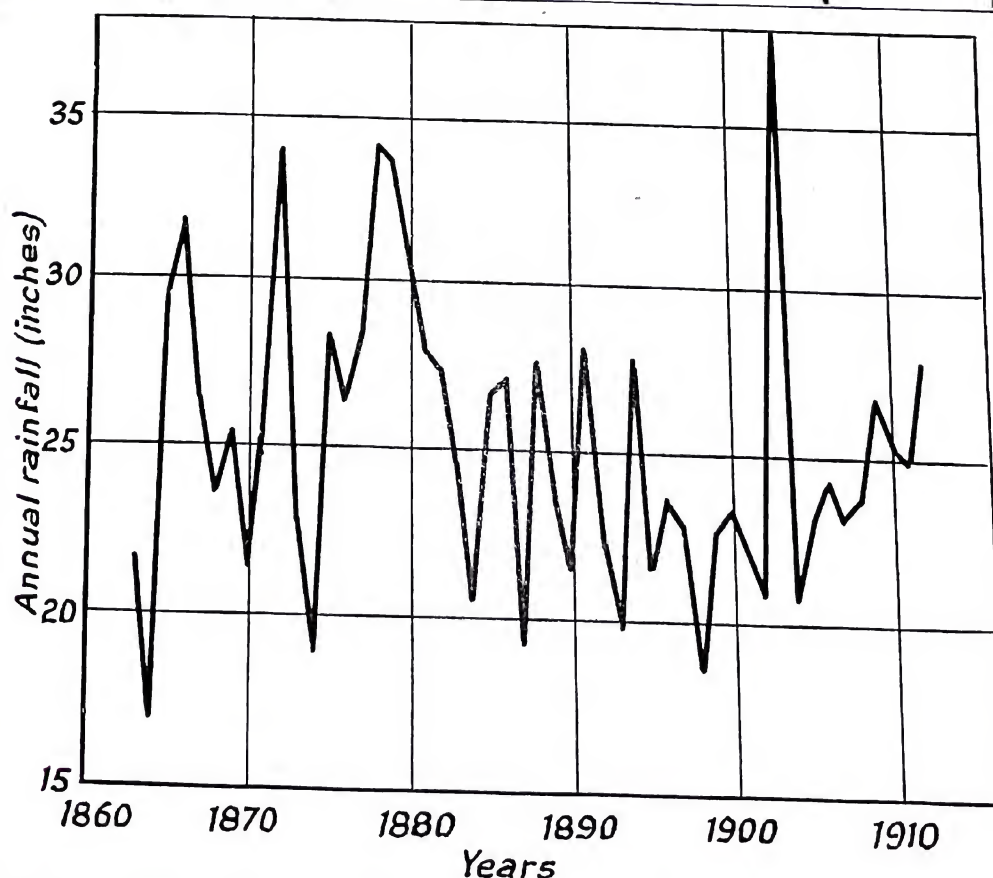


Fig. 45.2—Graph of the last 50 terms of the data of Table 45.2 (rainfall)

or may not embody an element of random variation, but in the majority of cases with which we shall be concerned some such element is present, if only as an error of observation. We may regard a set of values $u(t_1), u(t_2), \dots, u(t_n)$ as the observed values of a multivariate complex. Their characteristic feature, however, is that the order of the set t_1, t_2, \dots, t_n is material and not, for example, accidental as it would be for a random sample x_1, x_2, \dots, x_n , in which the suffixes are adjoined for convenience of identification.

Table 45.3—Average number of eggs per laying hen in the U.S.A. for each month of the years 1938–1940

(Data from Report of the Bureau of Agricultural Economics, U.S. Dept. of Agriculture, on the Poultry and Egg Situation, March 1941)

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1938	7.9	9.9	15.4	17.5	17.3	14.9	13.6	11.8	9.4	7.5	5.9	6.4
1939	8.0	9.7	14.9	17.0	17.0	14.6	13.2	11.7	9.3	7.4	6.0	6.8
1940	7.2	9.0	14.4	16.5	17.0	14.8	13.4	11.8	9.7	7.9	6.2	6.8

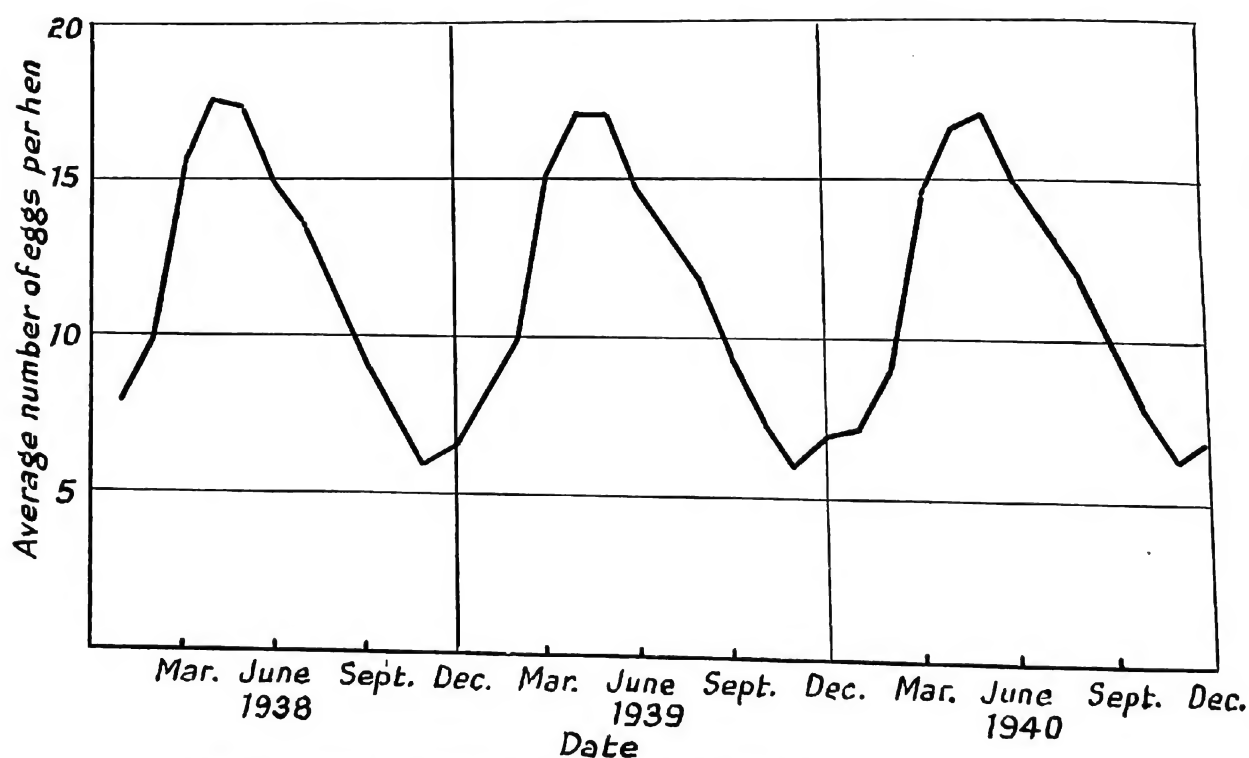


Fig. 45.3—Graph of the data of Table 45.3 (egg production)

45.2 Although the variable t will always be spoken of and thought of as a time-parameter, the theory we are about to develop has obvious applications to variation in space. For example, if we consider the variation in thickness of a cotton thread along its length l , or the variation in intensity of wire-worm infestation along a traverse d in a field, the variables l and d may be interpreted in a manner analogous to t . Indeed, it is possible to generalize to space the notion of a random variable dependent on more than one such parameter. This is not a generalization we shall attempt.

Table 45.4—Sheep population of England and Wales for each year from 1867 to 1939
(Data from the Agricultural Statistics)

Year	Population (10,000)	Year	Population (10,000)	Year	Population (10,000)	Year	Population (10,000)
1867	2203	1886	1892	1905	1823	1924	1484
68	2360	87	1919	06	1843	25	1597
69	2254	88	1853	07	1880	26	1686
70	2165	89	1868	08	1968	27	1707
71	2024	90	1991	09	2029	28	1640
72	2078	91	2111	10	1996	29	1611
73	2214	92	2119	11	1933	30	1632
74	2292	93	1991	12	1805	31	1775
75	2207	94	1859	13	1713	32	1850
76	2119	95	1856	14	1726	33	1809
77	2119	96	1924	15	1752	34	1653
78	2137	97	1892	16	1795	35	1648
79	2132	98	1916	17	1717	36	1665
80	1955	99	1968	18	1648	37	1627
81	1785	1900	1928	19	1512	38	1791
82	1747	01	1898	20	1338	39	1797
83	1818	02	1850	21	1383		
84	1909	03	1841	22	1344		
85	1958	04	1824	23	1384		

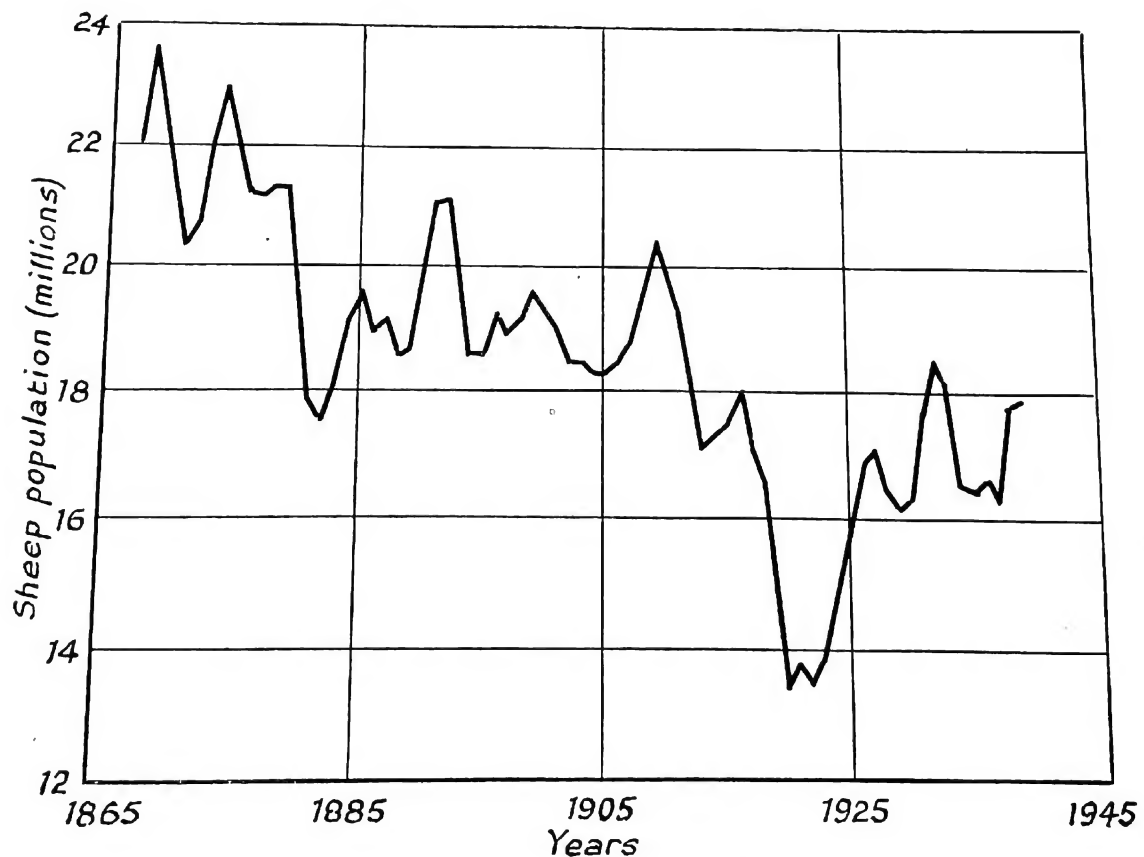


Fig. 45.4—Graph of the data of Table 45.4 (sheep population)

45.3 A further feature which distinguishes the set $u(t)$ from the values of a multivariate complex is that t is continuous, and we may therefore have to consider an infinity of values of $u(t)$. It is customary and convenient (though not, perhaps, very exact) to speak of a *continuous* time-series when we mean that t is continuous, not

Table 45.5—Population of England and Wales at ten-yearly intervals from 1811 to 1961
(Data from the Registrar-General's Statistical Review)

Year	Population (millions)	Year	Population (millions)
1811	10.16	1891	29.00
21	12.00	1901	32.53
31	13.90	11	36.07
41	15.91	21	37.89
51	17.93	31	39.95
61	20.07	41	—
71	22.71	51	43.76
81	25.97	61	46.07

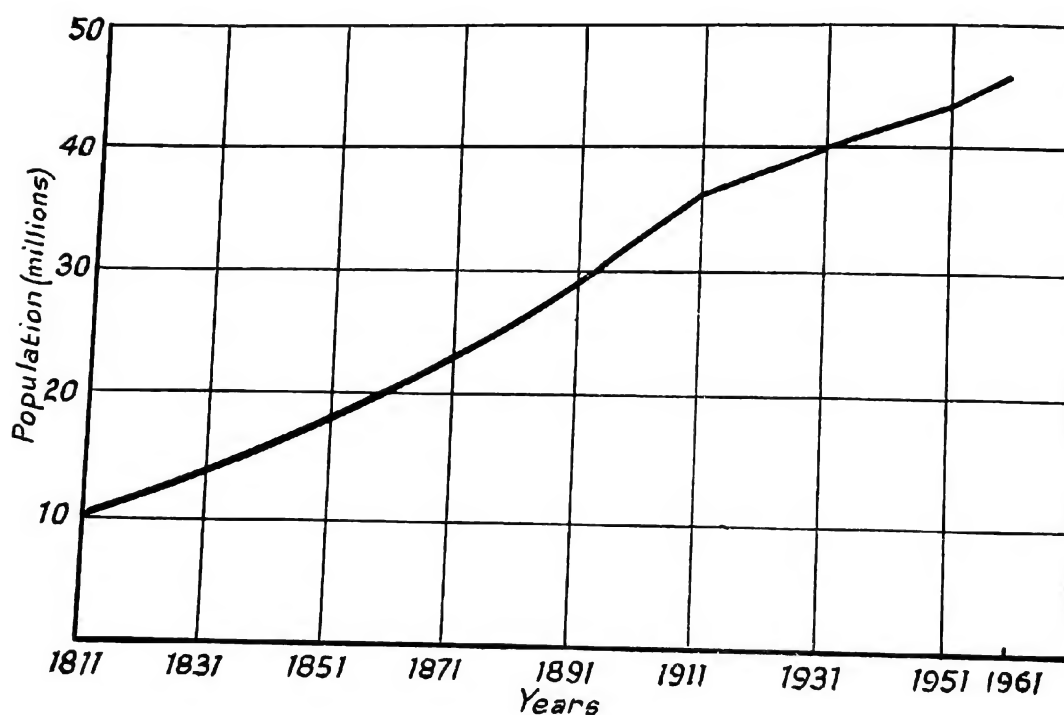


Fig. 45.5—Graph of the data of Table 45.5 (population of England and Wales)

necessarily implying that $u(t)$ is continuous for any given t in the variables under discussion. Likewise, by a *discontinuous* series we mean one given at a (discontinuous) set of points t_1, t_2, \dots, t_n , although u itself may be a continuous variable such as a length or a weight.

For the greater part of our treatment we shall be concerned with discontinuous series, but shall indicate applications to the continuous case where necessary. In fact, we shall mostly deal with series which are defined at equidistant points of time; and, taking the time-interval as unit, we may denote the values by u_0, u_1, u_2 , etc. If we

require to discuss values for time-points prior to the starting point u_0 , we may similarly denote them by u_{-1} , u_{-2} , and so on.

Some examples of time-series

45.4 The reader is doubtless familiar with many examples of time-series such as occur in ordinary life: the sales curve of a commodity over a period of years, the records of temperature or barometric pressure at a locality, drawn by a stylus on a rotating drum, the population of a country at a series of census dates, and so forth. We proceed to give a few specific examples which will indicate the kind of domain to be covered and serve for numerical exemplification of the theory to be developed later.

Table 45.1 (illustrated in Fig. 45.1) gives the annual yields per acre of barley in England and Wales from 1884 to 1939. Table 45.2 (Fig. 45.2) gives the annual rainfall in London for each year from 1813 to 1912. Table 45.3 (Fig. 45.3) gives the average egg-production per laying hen in the U.S.A. for each month of the years 1938 to 1940. Table 45.4 (Fig. 45.4) gives the sheep population of England and Wales as at June 4th of each year from 1867 to 1939. Table 45.5 (Fig. 45.5) shows the human population of England and Wales at 10-yearly intervals from 1811 to 1961.

45.5 These series are fairly typical of the kind of material with which our theory has to deal. The data of Table 45.1 (barley) present a very irregular fluctuation but, so far as the eye can see, there is no systematic element and no tendency towards increase or decrease over the period given. Table 45.2 has some indications of oscillatory movements of a more regular kind. Table 45.3 provides an oscillatory effect which is definitely seasonal. Table 45.4 (sheep population) combines a general decline in numbers with marked oscillatory effects. Table 45.5 (human population) shows a regular growth without apparent fluctuation.

Types of discontinuity

45.6 The tables also illustrate various types of discontinuity to which observed series are subject:

- (a) In the barley series we have a case of essential discontinuity. There is one and only one yield per acre for each year. The actual time of harvest may vary from year to year but, roughly speaking, the intervals between successive observations are equal.
- (b) In the population series we have a discontinuity of observation, due to the fact that a census is taken only every ten years. The variable, however, exists all through the period covered and could be observed (theoretically) at any point of time. The same is true of the sheep series.
- (c) In the rainfall data we have a discontinuity due to aggregation. The "rainfall" does not exist at a single point of time; it is the summation over a finite time-interval which is of interest. That interval, of course, is at choice. We may observe by year, by month, by day or even by hour. Intervals may overlap, as when we compile in successive weeks the rainfall for the previous month. Such data are nevertheless discontinuous time-series in our sense.

- (d) When we have a continuous record, e.g. on a barograph, we cannot tabulate it after the manner of Tables 45.1–45.5. We can take readings where we like, but not everywhere. In consequence, we cannot analyse such series by digital computation except as an approximation; we can, however, analyse them by methods involving graphical integration, e.g. by a planimeter or some more elaborate device.

45.7 In practice the time-points at which we observe the series are often determined for us, especially in economics. In experimental situations we may be able to decide them ourselves before the data are collected, or afterwards if a full record has been kept. The question what is the best interval of observation is to be decided in the light of the circumstances of the individual case, and is not one on which we can enter at this point. (Very little theoretical work has, in fact, been done on it.) We may note here, however, that observation at fixed equal intervals, convenient as it may be, can suppress evidence of oscillatory movements which have a period equal to those intervals or some sub-multiple of them. The annual observation of the sheep population, for example, will take no account of seasonal variation within the year due to slaughtering or breeding; the annual rainfall figures conceal the fact that rainfall is seasonal to some extent, even in London.

Calendar trouble

45.8 Whether time-intervals are equal or not, it is obviously desirable that observations should be comparable *inter se*. For series which are based on days or months there are certain nuisance-effects, due to the nature of the calendar, which have to be removed to ensure comparability. Some of these difficulties we can lay at the door of Nature, for not arranging that the year shall contain an integral number of days; but most of them are attributable to the man-made calendar. Months, for example, are not the same length; public holidays affect the comparability of economic and social data; exchanges and markets close over the week-end; and so on. Experimentally generated series are usually free from such difficulties if due care is taken, but they can arise in industrial series both in the large (e.g. stoppages due to strikes) or in the small (e.g. meal-breaks). We shall suppose that our data have been corrected for such effects so as to bring them on to a comparable basis.

The problems of time-series analysis

45.9 The ultimate object of analysis of a time-series—as of statistical analysis as a whole—is to arrive at a deeper understanding of the causal mechanisms which generated it, either out of sheer curiosity or because we wish to extrapolate into the future. It does not follow, however, that such understanding can be achieved by considering one series alone; for the series may be only a single facet of a complex phenomenon generating a substantial number of different series. We shall revert to this question in Chapter 50 when we discuss multivariate systems. For the present we curb our ambition to some extent by confining ourselves to the study of the type of behaviour of a single series and the setting up of models which can generate it; recognizing that such models themselves may be only portions of a more basic structural system. We shall see later that no logical inconsistency need be produced by this approach.

45.10 A survey of the practical examples we have given and of others known to the reader suggests that the typical time-series may be composed of four parts:

- (a) a trend, or long-term movement;
- (b) oscillations about the trend, of greater or less regularity;
- (c) a seasonal effect;
- (d) a "random," "unsystematic" or "irregular" component.

As a matter of mathematical description, we can always represent a series as one of these constituents or the sum of several of them. A large part of the traditional theory of time-series, in fact, is devoted to an analysis of the data into such components, so as to isolate them for separate study. We must, however, attempt to avoid a trap here. It does not follow that if we can represent a series as a sum of such components, they correspond to independently operating causal systems. The decomposition of a series is very often useful, but it may be misleading and in any case is not the ultimate object of statistical analysis.

45.11 Perhaps the easiest component to understand and to remove from the series is the seasonal effect. This is a fluctuation imposed on the series by a cyclic phenomenon external to the main body of causal influences at work upon it. The oscillation in egg-production in Table 45.3, for instance, reflects the rhythm in the reproductive process which is found among birds in virtue, ultimately, of the fact that the earth goes round the sun once a year. We shall confine the word "seasonal" to those effects which are annual in period; but the same ideas can be applied to any phenomenon generated by strictly periodic natural processes, such as "spring" and "neap" variation in tides or daily variation in temperature. We must, however, be careful about extending the notion of seasonality to phenomena which are not demonstrated beyond reasonable doubt to depend on strictly periodic stimuli. For instance, it would be going too far, in the present state of our knowledge, to speak of sunspot variation as seasonal in this sense, and much too far to speak of seasonality in crop-yields as determined by sunspots, even if the relation between the two were established. We shall return to this point below when defining what we mean by a "cycle" as distinct from an "oscillation."

45.12 The concept of trend is more difficult to define. Generally, one thinks of it as a smooth broad motion of the system over a long term of years, but "long" in this connexion is a relative term, and what is long for one purpose may be short for another. For example, if we were examining rainfall records over a hundred years, a slow rise from the beginning of the period to the end would be regarded as a trend; but if we possessed records for two thousand years (and the rings in some of the giant redwood trees give an index of climatic conditions for periods of this order) the rise over a particular century might appear as part of a slow oscillatory movement, so that any inference from the "trend" in a particular century to the effect that the weather was likely to continue becoming wetter and wetter might be quite false. What inference we should make in practice would depend on what we were trying to do. If we were engineers designing a water-supply system and wished to provide against droughts of

reasonable extent, we might perhaps assume that the trend would last as long as our works and proceed accordingly; but if we were attempting to study climatic changes over the face of the earth for geological periods of time we should accept the continuance of the trend with the greatest reserve or, more probably, should reject it on collateral grounds.

45.13 However long a series may be, we can never be certain, and often not even reasonably sure, that a trend in it is not part of a slow oscillation, except of course when the series has terminated (as might, for instance, be the case if we were considering the lengths of reigns of the Roman Emperors). In speaking of a trend, therefore, we must bear in mind the length of the series to which our statement refers. Perhaps it would be more accurate to speak of slow or quick movements rather than of trend and oscillation, but even so the distinction between the two would remain a matter of subjective judgement to some extent.

45.14 When seasonal variation and trend have been removed from the data we are left with a series which will present, in general, fluctuations of a more or less regular kind. Fig. 45.1 represents the kind of series we obtain, since it has no components of trend or seasonality. The question then arises, is this residual series systematic in the sense that its values can be represented as a function of the time? Or, on the other hand, are the values random in the sense that they could occur, in the observed order, by random sampling from a homogeneous population? Or again, is there some possibility intermediate between complete functional variation and complete randomness? The search for systematic effects in residual fluctuation gives rise to several techniques of analysis, the object of which is to detect whether any part of the series is subject to law, and therefore predictable, and whether any part is purely haphazard. The former part we shall call systematic, and it will be referred to as an "oscillation" (not a "cycle," which is a very special case of an oscillation, as we shall see later). The remainder of the series we shall call the unsystematic component, and refer to its movements as "random" or "stochastic." When a series is a mixture of oscillation and random movement it will not cause any inconvenience to refer to the up-and-down movement generally as fluctuation before we have analysed it into its constituents; that is to say, we may speak of fluctuation without prejudice to the possibility of detecting oscillatory movements in it.

Tests of randomness

45.15 Some of the series with which we are concerned are clearly not random. It would be a waste of time to test the data of Tables 45.3 and 45.5 for the presence of some systematic effects. In some cases, however, it is not obvious whether systematization is present, as for example in Table 45.1 (barley yields) and Table 45.2 (rainfall). We shall spend most of the rest of this chapter discussing tests of randomness in series. Specifically, given an ordered series of observations u_1, u_2, \dots, u_n , can they have arisen by chance in that order by sampling independently on n occasions from a population of unknown characteristics?

45.16 There is no limit to the number of tests which can be set up for this purpose. In choosing the most suitable we must have regard to a number of criteria:

- (a) If possible, the test should be distribution-free.
- (b) Since we may wish to test fairly long series, the calculations should be kept to a minimum.
- (c) Although we may not be able to specify an alternative hypothesis with precision, we may have some idea of its nature and can select a test which is likely to have high power against the alternative. For example, if we suspect trend we may find it useful to employ a different test from one used to test against periodicity.

We proceed to consider some tests satisfying these criteria.

Turning points

45.17 One of the easiest tests to apply is to count the number of peaks or troughs in the series. A "peak" is a value which is greater than the two neighbouring values. If there are two or more equal values which are greater than their predecessor and successor (a rare event in general) we shall regard them as defining one peak. Likewise a "trough" is a value which is lower than its two neighbours. Our first question is: What is the distribution of peaks in a random series? (The distribution of troughs is evidently the same with a change of sign of the variate.)

In point of fact, we shall find it more convenient to treat both peaks and troughs as cases of "turning points" of the series. The number of turning points is clearly one less than the number of runs up and down in the series. The interval between two turning points is called a "phase."

45.18 Three consecutive observations are required to define a turning point, say u_1, u_2, u_3 . If the series is random these three values could have occurred in any order, namely in six ways. In only four of these ways would there be a turning point (when the greatest or least value is in the middle). Hence the probability of a turning point in a set of three values is $\frac{2}{3}$.

Consider now a set of values u_1, u_2, \dots, u_n , and let us define a "marker" variable X_i by

$$\begin{aligned} X_i &= 1, & u_i < u_{i+1} > u_{i+2} \\ & \text{or} & u_i > u_{i+1} < u_{i+2} \\ &= 0 & \text{otherwise; } i = 1, 2, \dots, n-2. \end{aligned} \quad (45.1)$$

The number of turning points p is then simply

$$p = \sum_{i=1}^{n-2} X_i. \quad (45.2)$$

We have at once

$$E(p) = \sum E(X_i) = \frac{2}{3}(n-2). \quad (45.3)$$

Also $E(p^2) = E\left(\sum_1^{n-2} X_i\right)^2$

$$= E\left\{ \sum_{i=1}^{n-2} X_i^2 + 2 \sum_{i=1}^{n-3} X_i X_{i+1} + 2 \sum_{i=1}^{n-4} X_i X_{i+2} + \sum_{\substack{(n-4)(n-5) \\ k \neq 0, 1, 2}} X_i X_{i+k} \right\}, \quad (45.4)$$

where the suffixes to the Σ signs indicate the number of terms over which summation takes place. As a check note that

$$(n-2)^2 = (n-2) + 2(n-3) + 2(n-4) + (n-4)(n-5).$$

We then have

$$E(p^2) = (n-2)EX_i^2 + 2(n-3)E(X_i X_{i+1}) + 2(n-4)E(X_i X_{i+2}) + (n-4)(n-5)E(X_i X_{i+k}). \quad (45.5)$$

Since $X_i^2 = X_i$ we have

$$E(X_i^2) = \frac{2}{3}. \quad (45.6)$$

For $k > 2$, X_i and X_k are independent, for they have no value of u in common. Thus

$$E(X_i X_{i+k}) = E(X_i)E(X_{i+k}) = \frac{4}{9}. \quad (45.7)$$

It remains to evaluate $E(X_i X_{i+1})$ and $E(X_i X_{i+2})$. For the first, consider four consecutive terms which, in ascending order of magnitude, may be denoted by the numbers 1, 2, 3, 4. The only non-vanishing contribution to $X_i X_{i+1}$ which can arise from a permutation of these numbers arises when there is a turning point in the second and third places. If the reader will write down the 24 possible permutations he will find that only ten make a non-vanishing contribution, namely

1324	2143	3142	4132
1423	2314	3241	4231
	2413	3412	

Thus

$$E(X_i X_{i+1}) = \frac{10}{24} = \frac{5}{12}. \quad (45.8)$$

For $X_i X_{i+2}$ we have to write down the 120 permutations of the integers 1 to 5 and count up those with turning points at both the second and fourth three places. There are, in fact, 54 and thus

$$E(X_i X_{i+2}) = \frac{54}{120} = \frac{9}{20}. \quad (45.9)$$

Substituting in (45.5) we find, on reduction,

$$E(p^2) = \frac{40n^2 - 144n + 131}{90}.$$

Hence, using (45.3), we have

$$\text{var } p = \frac{16n - 29}{90}. \quad (45.10)$$

Higher moments can be obtained in a similar manner. We find

$$\kappa_3(p) = \frac{-16(n+1)}{945} \quad (45.11)$$

$$\kappa_4(p) = \frac{-1408n + 3317}{18900}. \quad (45.12)$$

Thus, in standard measure $\kappa_3(p)$ is approximately $0.2n^{-1}$ and $\kappa_4(p)$ is approximately $1.2n^{-1}$, indicating a fairly rapid tendency to normality as n increases.

45.19 Now consider the distribution of phase lengths. To define a phase of length d (say, a run up) we require $d+3$ terms, involving a fall from first to second, a rise from second to third, third to fourth, $(d+1)$ th to $(d+2)$ th, and a fall from $(d+2)$ th to $(d+3)$ th. Consider the $d+3$ values arranged in increasing order of magnitude. If we pick out two other than the first and the last and transfer one to the beginning and one to the end, we obtain a rising phase of length d . There are $\frac{1}{2}(d+1)d$ ways of picking out the pair, and each may go to either end, so there are $(d+1)d$ rising phases. But in addition we may put the first member at the end and any of the others except the second at the beginning, giving us $d+1$ further cases; or the last member at the beginning and any except the penultimate at the end, giving $(d+1)$ further cases; and from this total we must subtract the case where the first is last and the last first, which has been counted twice. Thus there are

$$(d+1)d + (d+1) + (d+1) - 1 = d^2 + 3d + 1$$

rising phases. The probability of a phase, either rising or falling, is then

$$\frac{2(d^2 + 3d + 1)}{(d+3)!}. \quad (45.13)$$

Now in a series of length n there are $n-d-2$ possible phases of length d . The expected number of phases of length d in the set of n values is then

$$N_d = \frac{2(n-d-2)(d^2 + 3d + 1)}{(d+3)!}. \quad (45.14)$$

The expected total number of phases N , from (45.14), is given by

$$N = 2 \sum_{d=1}^{n-3} \frac{(n-d-2)(d^2 + 3d + 1)}{(d+3)!}.$$

$$\text{Now} \quad (n-d-2)(d^2 + 3d + 1) = -(d+3)(d+2)(d+1) + (n+1)(d+3)(d+2) \\ - (2n+1)(d+3) + (n+1)$$

and hence

$$N = 2 \sum_{d=1}^{n-3} \left\{ -\frac{1}{d!} + \frac{n+1}{(d+1)!} - \frac{2n+1}{(d+2)!} + \frac{n+1}{(d+3)!} \right\} \\ = 2 \left(\frac{2n-7}{6} + \frac{1}{n!} \right). \quad (45.15)$$

For all practical purposes we may neglect the second factor in (45.15) and hence

$$N \doteq \frac{1}{3}(2n-7). \quad (45.16)$$

Since the number of phases is one less than the number of turning points except in the 2 cases out of $n!$ where both are zero, (45.15) agrees with (45.3). Now

$$N_d/N \doteq \frac{6(n-d-2)(d^2 + 3d + 1)}{(d+3)!(2n-7)}. \quad (45.17)$$

We may derive the moments of this ratio fairly easily. For example,

$$\mu'_1 = \frac{6}{2n-7} \sum_1^{n-3} \frac{d(n-d-2)(d^2 + 3d + 1)}{(d+3)!}$$

$$= \frac{6}{2n-7} \sum_1^{n-3} \left\{ -\frac{1}{(d-1)!} + \frac{n+1}{d!} - \frac{3n+2}{(d+1)!} + \frac{5n+3}{(d+2)!} - \frac{3(n+1)}{(d+3)!} \right\}.$$

Remembering the rapid convergence of $\sum_0^n 1/x!$ to e , we may to a very close approximation write this as

$$\begin{aligned} \mu'_1 &= \frac{6}{2n-7} \left\{ -e + (n+1)(e-1) - (3n+2)(e-2) + (5n+3)\left(e-\frac{5}{2}\right) - 3(n+1)\left(e-\frac{8}{3}\right) \right\} \\ &= \frac{3(n+7-4e)}{2n-7} \doteq \frac{3}{2}. \end{aligned} \quad (45.18)$$

Likewise we find that

$$\mu_2 = \frac{3}{(2n-7)^2} \{(8e-21)n^2 + (4e-17)n - (48e^2 - 140e + 14)\} \doteq 0.560. \quad (45.19)$$

45.20 The distribution of which these are the moments does not tend to normality as n increases (cf. Exercise 45.1). A natural procedure in testing for randomness is to compare the observed distribution with the expected distribution given by (45.14). For shorter series, however, there is a theoretical difficulty in that the lengths of phase are not independent, so that a straightforward χ^2 goodness-of-fit test is not valid. The question was examined by Wallis and Moore (1941) who came to the conclusion that for a three-fold classification $d = 1, 2, \geq 3$ (two degrees of freedom) the X^2 statistic(*) can be tested in the ordinary form with $\nu = 2\frac{1}{2}$ for $X^2 \geq 6.3$. For lower values $\frac{6}{7}X^2$ can be tested in that form with $\nu = 2$.

Wolfowitz (1944) and Levene (1952) showed that the number of phases tends to normality and Gleissberg (1945) tabulated the distribution of this number for $n \leq 25$.

Example 45.1

Consider the barley yields of Table 45.1. There are 56 values in this series, but at two points (1906, 1907 and 1910, 1911) the values in successive years are equal. So far as concerns turning points and phases we shall count each of these as one point and reduce the number of terms to 54.

If the reader will mark the peaks and troughs on the table, or count them on Fig. 45.1, he will find that there are 35 turning points. The expected number, from (45.3), is $\frac{2}{3}(52) = 34\frac{2}{3}$. This is so close to observation that no further test is necessary.

The distribution of phases will be found to be

Phase length	No. of phases observed	No. of phases, theoretical (45.14), (45.16)
1	23	21.25
2	7	9.17
3	4	2.59
TOTAL	34	33.67

Again a test is hardly necessary.

(*) See footnote to page 421, Vol. 2.

The conclusion would be, on these tests, that the variation in yield from year to year was random.

45.21 Considered as a test against trend the turning-points test has a poor performance, and we shall see later (Exercise 45.4) that it has zero efficiency compared with other tests in certain cases. This is intuitively reasonable, for "turning" is a local property and would not be much affected by whereabouts along a line of gentle trend development the series had arrived. Considered as a test against cyclicity the test is obviously better. In a random series the mean interval between turning points is about 1.5 with a variance (from (45.10) and (10.14)) of about $9/(10n)$. The test itself is enough to enjoin further investigation in series of more than 10 terms whenever the mean interval between turning points is 2 or more.

The power of tests against specific alternatives for runs up and down has been investigated by Levene (1952).

The difference-sign test

45.22 A somewhat more laborious test consists of counting the number of positive first differences of the series, that is to say, the number of points where the series increases. (As before, we shall ignore points where there is neither increase nor decrease.) With a series of n terms we have $n-1$ differences. Let us, as before, define a variable

$$\begin{aligned} X_i &= 1, & u_{i+1} > u_i \\ &= 0 & u_{i+1} < u_i; \quad i = 1, 2, \dots, (n-1). \end{aligned} \quad (45.20)$$

Then the number of points of increase, say c , is given by

$$c = \sum_{i=1}^{n-1} X_i.$$

For a random series we have immediately

$$E(c) = (n-1)E(X_i) = \frac{1}{2}(n-1). \quad (45.21)$$

Likewise

$$\begin{aligned} E(c^2) &= E\left\{ \sum_{i=1}^{n-1} X_i^2 + 2 \sum_{i=2}^{n-2} X_i X_{i+1} + \sum_{(i-2)(i-3)}^{(n-2)(n-3)} X_i X_j \right\}, \quad j \neq i, i+1, \\ &= (n-1)EX_i^2 + 2(n-2)E(X_i X_{i+1}) + (n-2)(n-3)E(X_i X_j). \\ &= \frac{1}{2}(n-1) + 2(n-2)E(X_i X_{i+1}) + \frac{1}{4}(n-2)(n-3). \end{aligned} \quad (45.22)$$

To evaluate $E(X_i X_{i+1})$ we consider permutations of three. Only in one case out of six does this give a non-vanishing contribution. Hence, from (45.22) and (45.21) we find

$$\begin{aligned} \text{var } c &= \frac{1}{2}(n-1) + \frac{1}{3}(n-2) + \frac{1}{4}(n-2)(n-3) - \frac{1}{4}(n-1)^2 \\ &= \frac{1}{12}(n+1). \end{aligned} \quad (45.23)$$

The distribution tends fairly rapidly to normality (cf. Exercise 45.3). It has been tabulated by Moore and Wallis (1943).

45.23 This test is clearly useless against an alternative of symmetrical oscillation, where the number of movements up will approximate to the number of movements

down. It has been advocated mainly as a test against trend, and especially against linear trend. As such it is very superior to the turning-points test but very inferior to other tests based on rank order which we consider below.

Consider, in fact, a series with a linear trend and a random residual

$$u_t = \alpha + \beta t + \varepsilon_t \quad (45.24)$$

where ε_t is a normal variable with zero mean and unit variance. We can regard this as a regression of u_t on t in the special case when t takes the equidistant values $1, 2, \dots, n$. In the regression situation we should estimate β by

$$b = \frac{\sum (u_t - \bar{u})(t - \bar{t})}{\sum (t - \bar{t})^2} \quad (45.25)$$

which is unbiased and has variance

$$\text{var } b = \frac{1}{\sum (t - \bar{t})^2} = \frac{12}{n(n^2 - 1)} \doteq \frac{12}{n^3}. \quad (45.26)$$

We now use the asymptotic relative efficiency (cf. 25.5-6) to compare other consistent tests with that based on b . Since b is unbiased, $[\partial E(b)/\partial \beta]_{\beta=0} = 1$, and (25.16) becomes, with $m = 1$,

$$[E'(b)]_{\beta=0}/(\text{var } b)^{\frac{1}{2}} \sim (n^3/12)^{\frac{1}{2}}. \quad (45.27)$$

Thus, for the statistic b , δ defined by (25.16) takes the value

$$\delta_b = \frac{3}{2}. \quad (45.28)$$

We now compute δ for the difference-sign test statistic.

Consider the "marker" variable

$$\begin{aligned} H_{ij} &= 1, & u_i > u_j \\ &= 0 & u_i < u_j \end{aligned} \quad (45.29)$$

with

$$H_{ii} = 1.$$

The expectation of H_{ij} is the probability that H_{ij} equals unity, and since $u_i - u_j$ is a normal variable with mean $\beta(i-j)$ and variance 2, this is equal to

$$\begin{aligned} &\int_0^\infty \frac{1}{\sqrt{(2\pi)2^{\frac{1}{2}}}} \exp \left[-\frac{1}{4} \{x - \beta(i-j)\}^2 \right] dx \\ &= \frac{1}{\sqrt{(2\pi)}} \int_{-\frac{1}{\sqrt{2}}\beta(i-j)}^\infty \exp \left(-\frac{1}{2} y^2 \right) dy. \end{aligned}$$

Hence

$$\left[\frac{\partial}{\partial \beta} E(H_{ij}) \right]_{\beta=0} = \frac{i-j}{\sqrt{2}} \cdot \frac{1}{\sqrt{(2\pi)}} = \frac{i-j}{2\sqrt{\pi}}. \quad (45.30)$$

This is all we require for the ARE of the present test, but for later purposes we proceed to calculate some other quantities of a similar kind.

Since H_{ij} , H_{kl} are independent, i, j, k, l unequal, we have

$$\begin{aligned} \left[\frac{\partial}{\partial \beta} E(H_{ij} H_{kl}) \right]_{\beta=0} &= \left[E(H_{ij}) \frac{\partial}{\partial \beta} E(H_{kl}) \right]_{\beta=0} + \left[E(H_{kl}) \frac{\partial}{\partial \beta} E(H_{ij}) \right]_{\beta=0} \\ &= \frac{1}{4\sqrt{\pi}} \{i-j+k-l\}. \end{aligned} \quad (45.31)$$

Consider now $H_{ij}H_{jk}$. Since $y_i - y_j$ and $y_j - y_k$ are jointly normally distributed with correlation $-\frac{1}{2}$ we have

$$E(H_{ij}H_{jk}) = \text{Prob}(H_{ij} = 1, H_{jk} = 1) \\ = \int_{-\beta(i-j)/\sqrt{2}}^{\infty} \int_{-\beta(j-k)/\sqrt{2}}^{\infty} \frac{1}{2n\sqrt{\frac{3}{4}}} \exp\left\{-\frac{2}{3}(x^2 - xy + y^2)\right\} dx dy$$

and

$$\left[\frac{\partial}{\partial\beta}(H_{ij}H_{jk})\right]_{\beta=0} = \frac{i-j}{\sqrt{2}} + \frac{j-k}{\sqrt{2}} \int_0^{\infty} \frac{1}{2\pi} \exp -\frac{1}{2}x^2 dx \\ = \frac{i-k}{4\sqrt{\pi}}. \quad (45.32)$$

Similarly

$$\left[\frac{\partial}{\partial\beta}E(H_{ik}H_{jk})\right]_{\beta=0} = \frac{i+j-2k}{4\sqrt{\pi}}. \quad (45.33)$$

Finally we require the similar expression when $y_{i+2} - y_{i+1}$ and $y_{i+1} - y_i$ are of opposite signs. This is the probability that $(y_{i+2} - y_{i+1})(y_{i+1} - y_i)$ is negative and is seen to be

$$E = \int_{-\infty}^{-\beta/\sqrt{2}} \int_{-\beta/\sqrt{2}}^{\infty} \frac{1}{2\pi\sqrt{\frac{3}{4}}} \exp\left\{-\frac{2}{3}(x^2 - xy + y^2)\right\} dx dy$$

whence we find

$$\left[\frac{\partial E}{\partial\beta}\right]_{\beta=0} = 0. \quad (45.34)$$

Reverting to the difference-sign test, we have that the number of positive increments in the series—cf. (45.20)—is

$$c = \sum_{i=1}^{n-1} H_{i, i+1}.$$

Thus from (45.30),

$$\left[\frac{\partial E(c)}{\partial\beta}\right]_{\beta=0} = \sum_{i=1}^{n-1} \frac{1}{2\sqrt{\pi}} = \frac{n-1}{2\sqrt{\pi}}. \quad (45.35)$$

Remembering that the variance is, from (45.23), $(n+1)/12$, we have

$$\left(\frac{\partial E(c)}{\partial\beta}\right)_{\beta=0} / (\text{var } c)^{\frac{1}{2}} \sim \left(\frac{3n}{\pi}\right)^{\frac{1}{2}}. \quad (45.36)$$

Thus, for the difference-sign test, δ defined at (25.16), Vol. 2, takes the value

$$\delta_c = \frac{1}{2}, \quad (45.37)$$

and comparison of (45.37) and (45.28) shows that c has zero asymptotic relative efficiency, by (25.24).

Mann (1945a) gave a lower bound to the power of the test. Stuart (1952) has tabulated the power of the test against the normal regression alternative at the 95 per cent level.

Rank correlation tests

45.24 There is a prior presumption that we shall improve our test still further if we compare, not merely neighbouring pairs as in the difference-sign test, but all pairs. Given a set of values u_1, u_2, \dots, u_n , in that order, let us count the number

of pairs in which $u_j > u_i, j > i$. If this is P , we note that there are $\frac{1}{2}n(n-1)$ pairs and that the expected number in a random series is $\frac{1}{4}n(n-1)$. The excess of P over this number indicates a tendency to positive trend, a deficiency corresponding to a negative trend.

In fact, this quantity is a simple linear function of the rank correlation coefficient(*) τ , defined at (31.23), Vol. 2, between the order of the variables in time and their order in magnitude u . For a random series the variance of τ is known. If Q is the complementary quantity to P , namely the number of values for which $u_j < u_i, j > i$, we have, by (31.23),

$$\tau = 1 - \frac{4Q}{n(n-1)}, \quad (45.38)$$

and from (31.33-4),

$$E(\tau) = 0 \quad (45.39)$$

$$\text{var } \tau = \frac{2(2n+5)}{9n(n-1)}. \quad (45.40)$$

The distribution of τ tends rapidly to normality—cf. 31.26.

45.25 In the notation of 45.23 we may write

$$Q = \sum_{i < j}^n H_{ij}.$$

In the case of an alternative linear trend (45.24) with normal residuals, we then have, from (45.30),

$$\begin{aligned} \left[\frac{\partial}{\partial \beta} E(Q) \right]_{\beta=0} &= \frac{1}{2\sqrt{\pi}} \sum_{i < j}^n (i-j) \\ &= -\frac{1}{2\sqrt{\pi}} \sum_{j=1}^n \left\{ \frac{1}{2}j(j+1) - j \right\} \\ &= -\frac{1}{2\sqrt{\pi}} \cdot \frac{n(n^2-1)}{6}. \end{aligned}$$

Also, from (45.40),

$$\text{var } Q \sim \frac{n^3}{36},$$

so that
$$\left[\frac{\partial}{\partial \beta} E(Q) \right]_{\beta=0} / (\text{var } Q)^{\frac{1}{2}} \sim \{n^3/(4\pi)\}^{\frac{1}{2}}. \quad (45.41)$$

Substituting (45.41) and (45.27) in (25.27) we have for the ARE of τ (or Q) relative to the regression estimator,

$$A_{t,b} = \left\{ \frac{[E'(t)]_{\beta=0}/(\text{var } t)^{\frac{1}{2}}}{[E'(b)]_{\beta=0}/(\text{var } b)^{\frac{1}{2}}} \right\}^{\frac{1}{(3/2)}} = (3/\pi)^{\frac{1}{2}} \doteq 0.98, \quad (45.42)$$

a result previously given in 31.38. The statistic τ is therefore very efficient in this case.

Mann (1945b) considered the τ -test for variables u_i such that $P(u_i > u_j) = \frac{1}{2} + \varepsilon_{ij}$ for $i < j$, and obeying certain other conditions, and (cf. Exercise 31.8) gave conditions for the test to be unbiased, and an example in which it is the most powerful test.

(*) In other contexts we write t for the statistic τ , but as this would cause confusion here with the time-variable, we depart temporarily from the usual convention not to employ Greek letters for sample values.

45.26 We can also calculate the Spearman rank correlation r_s (31.22, Vol. 2) in this case. It may be written (cf. (31.40))

$$r_s = 1 - \frac{12V}{n(n^2-1)} \quad (45.43)$$

where

$$V = \sum_{i < j}^n (j-i)H_{ij}. \quad (45.44)$$

As at (31.22) we have, for a random series,

$$E(r_s) = 0$$

$$\text{var } r_s = \frac{1}{n-1},$$

so that

$$E(V) = \sum_{i < j} (j-i)E(H_{ij}) \quad (45.45)$$

$$\text{var } V \doteq \frac{n^5}{144}. \quad (45.46)$$

We then find

$$\begin{aligned} \left[\frac{\partial}{\partial \beta} E(V) \right]_{\beta=0} &= -\frac{1}{2\sqrt{\pi}} \sum_{i < j} (j-i)^2 \\ &= -\frac{n^2(n^2-1)}{24\sqrt{\pi}}, \end{aligned} \quad (45.47)$$

and exactly as at (45.42) we find

$$A_{r_s, b} = \left(\frac{3}{\pi} \right)^{\frac{1}{2}} = 0.98. \quad (45.48)$$

The Spearman coefficient has, then, the same ARE as τ .

45.27 Both τ and r_s are more troublesome to calculate than the difference-sign or the turning-point statistics; and in practice r_s is easier to calculate than τ , using its form (31.21).

Example 45.2

Not to overburden ourselves with arithmetic, let us take the first twenty-five terms in Table 45.2 (years 1813–1837). In order of magnitude the values of u_t are

Rank	Rank	Difference ²	Rank	Rank	Difference ²	Rank	Rank	Difference ²
1	9	64	10	11	1	19	21	4
2	18	256	11	13	4	20	2	324
3	4	1	12	25	169	21	15	36
4	23	361	13	8	25	22	3	361
5	10	25	14	5	81	23	14	81
6	12	36	15	7	64	24	20	16
7	19	144	16	22	36	25	1	576
8	6	4	17	17	0			2898
9	24	225	18	16	4			

Thus

$$V = \frac{1}{2}(2898) = 1449$$

$$r_s = 1 - \frac{(12)(1449)}{15600} = -0.114.$$

The correlation is small. The standard error of r_s is $(n-1)^{\frac{1}{2}}$, about 0.2, and the observed value is thus not significant of trend.

There are 17 turning points, against an expectation of $\frac{2}{3}(25) = 16.7$, in almost perfect agreement.

45.28 We may also mention, without entering into very great detail, two other tests which have some interest:

- (a) *The records test.* An observation u_i is called an upper (lower) record if it exceeds (is smaller than) all previous observations in the series. The number of records appearing as we go along the series provides a test statistic which can be compared with the distribution from a random series. The subject has been explored by Foster and Stuart (1954), some of whose results are presented in Exercises 45.8-9. It appears that, as a test against trend, the records test is more powerful than the difference-sign test or the turning-points test, but is considerably less powerful than the τ or r_s tests (Stuart, 1956, 1957).
- (b) *The rank serial correlation test.* This is a special case of a type of statistic, the serial correlation coefficient, which we shall introduce in 45.32 below. If the ranks of a set of n quantities measured about the mean $\frac{1}{2}(n+1)$ are d_i , $i = 1, 2, \dots, n$, the coefficient of order k is defined by

$$r_k = \frac{\frac{1}{n-k} \sum_{i=1}^{n-k} d_i d_{i+k}}{\frac{1}{12}(n^2-1)}. \quad (45.49)$$

So far as a test statistic is concerned we may use simply

$$W_k = \sum_{i=1}^{n-k} d_i d_{i+k}. \quad (45.50)$$

The coefficient W_k is the covariance (multiplied by $n-k$) of the terms in the rank-series distance k units apart. For a random series its expected value is zero. As a test of trend the coefficient has zero efficiency against the normal linear regression alternative (cf. Noether (1950), Stuart (1954, 1956)), although it was suggested as a test against trend by Wald and Wolfowitz (1943).

45.29 If a series is random we shall clearly be able to test it equally well by ignoring certain terms, e.g. by taking every other term or every twelfth term. To look at this from another viewpoint, if our series is only recorded at periodic intervals instead of *in toto*, our tests remain valid. We lose information, of course, but not validity in the tests. The same is true of aggregative series; for example, if we have the (annual) aggregate of twelve monthly records, each of which may be regarded as a member of a random series, the annual figures are also random. On the other hand, randomness in an annual series does not rule out the possibility of seasonal movements in the constituent series.

45.30 Some series (for example, a razor blade under the microscope) present an irregular fluctuation which looks like a kind of randomness in the limit as the interval between successive observations becomes smaller. On the other hand, the series are continuous, and we arrive at the question whether it is possible to have a continuous random series. In our view, it is not. There is, to our mind, something essentially discontinuous in the idea of independence of successive observations; continuity would destroy independence. We can imagine a set of points, each determining the value of a random variable, becoming ever closer together, and the variance of the variable diminishing so that the total range of variation remains within finite bounds. But it does not appear possible to proceed to the limit in the way that the mathematician proceeds from an enumerable to a dense set of points on a line.

45.31 For most, if not all, practical series we can imagine them as continuous to the eye but discontinuous under the microscope. Pressure may be thought of as a continuous variable but on a sufficiently small time-scale is discontinuous, being the result of impacts of individual molecules of gas. The profile of the cotton fibre may be similarly imagined as continuous, although ultimately composed of discontinuous particles. In this sense we may, perhaps, speak of a continuous random series, but it is a form of expression which we shall have to watch very carefully. To test such a series for randomness we may take observations at any suitable interval and carry out on the resultant one of the tests we have discussed earlier in the chapter. A discussion of more refined methods of approach must await our account of correlogram and spectral analyses.

Serial correlations

45.32 For series which are not random there will be dependencies of one kind or another between successive terms. One very useful measure of this effect is the product-moment correlation between successive observations. Given n values u_1, u_2, \dots, u_n , the so-called serial correlation of lag 1 is defined by

$$r_1 = \frac{\frac{1}{n-1} \sum_{i=1}^{n-1} \left\{ \left(u_i - \frac{1}{n-1} \sum_{i=1}^{n-1} u_i \right) \left(u_{i+1} - \frac{1}{n-1} \sum_{i=1}^{n-1} u_{i+1} \right) \right\}}{\left[\frac{1}{n-1} \sum_{i=1}^{n-1} \left\{ u_i - \frac{1}{n-1} \sum_{i=1}^{n-1} u_i \right\}^2 \right]^{\frac{1}{2}} \left[\frac{1}{n-1} \sum_{i=1}^{n-1} \left\{ u_{i+1} - \frac{1}{n-1} \sum_{i=1}^{n-1} u_{i+1} \right\}^2 \right]^{\frac{1}{2}}}. \quad (45.51)$$

Likewise, the serial correlation of lag k is the correlation between pairs of terms k units apart, viz.

$$r_k = \frac{\frac{1}{n-k} \sum_{i=1}^{n-k} \left(u_i - \frac{1}{n-k} \sum_{i=1}^{n-k} u_i \right) \left(u_{i+k} - \frac{1}{n-k} \sum_{i=1}^{n-k} u_{i+k} \right)}{\left[\frac{1}{n-k} \sum_{i=1}^{n-k} \left\{ u_i - \frac{1}{n-k} \sum_{i=1}^{n-k} u_i \right\}^2 \right]^{\frac{1}{2}} \left[\frac{1}{n-k} \sum_{i=1}^{n-k} \left\{ u_{i+k} - \frac{1}{n-k} \sum_{i=1}^{n-k} u_{i+k} \right\}^2 \right]^{\frac{1}{2}}}. \quad (45.52)$$

In practice (and also for theoretical convenience) it makes for simplicity to modify these definitions to some extent. Instead of measuring the first $(n-k)$ u 's about their mean, we may measure about the mean of the whole set of observations; and similarly for the values at the end. Similarly, instead of taking separate variances in the

denominator of (45.52) we may use the variance of the whole series. Thus, writing \bar{u} for $\sum_{i=1}^n u_i/n$, we may put

$$r_k = \frac{\frac{1}{n-k} \sum_{i=1}^{n-k} (u_i - \bar{u})(u_{i+k} - \bar{u})}{\frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^2}. \quad (45.53)$$

This is the form we shall mostly use. For series of moderate length the difference from (45.52) is negligible. We must be careful not to use (45.53) for short series where exactitude in estimation is necessary. In particular, values of r_k greater than unity may arise.

45.33 The array of coefficients $r_0 (=1), r_1, r_2, \dots$ tells us a good deal about the nature of the internal dependence of the series. Their totality is called the *correlogram*, a term which is also used to denote the graph of r_k as ordinate against k as abscissa. In a random series they are, apart from r_0 , all equal to zero within sampling limits. We shall study their properties for other types of series at length in later chapters. It is to be noted that, by definition, $r_{-k} = r_k$.

45.34 For certain theoretical enquiries, and for computational convenience in a minor way, the definition (45.53) may be modified still further. For a coefficient of order k there are only $n-k$ terms in the numerator. Suppose we put

$$u_{n+1} = u_1, u_{n+2} = u_2, \dots, u_{n+k} = u_k. \quad (45.54)$$

We may then sum the product-moment in the numerator over n terms to obtain

$$r_k = \frac{\sum_{i=1}^n (u_i - \bar{u})(u_{i+k} - \bar{u})}{\sum_{i=1}^n (u_i - \bar{u})^2}. \quad (45.55)$$

We are obviously here distorting the data by assuming (45.54). But if k is small compared to n we are not distorting it very much. The coefficient r_k of (45.55) is called a *circular* serial correlation. The point of introducing it will become evident in 47.30 and Chapter 48.

45.35 In concluding this chapter we may, for the avoidance of confusion, refer to a type of occurrence which is, in a sense, a time-series, though not of the kind we are considering here. Events such as accidents, arrivals of cars at traffic lights, outbreaks of epidemics, etc., may happen from time to time in a certain area and thus over a period constitute a series of events. The intervals between them are, usually, irregular but may nevertheless have a distribution function. Such patterns of behaviour are studied in the theory of stochastic processes. It is the intervals between happenings, rather than the happenings themselves, which are of interest, and this topic is really quite distinct from our time-series, which concerns a complex moving through time and observed at specified intervals.

EXERCISES

45.1 In the distribution of phases of 45.19 show that the moment-generating function of N_d/N for large n is given by

$$3(e^{-\theta} - 2e^{-2\theta} + e^{-3\theta})(g - 1 - e^\theta) + \frac{5}{2} - \frac{3}{2}e^{-\theta},$$

where

$$g = \exp(e^\theta).$$

Hence verify that $\mu'_1 = 1.5$, $\kappa_2 = 0.560$, and show that $\kappa_3 = 0.677$, $\kappa_4 = 0.904$. The distribution is thus not normal.

45.2 For the distribution of positive first differences (from a random series) of 45.22 show that odd order cumulants vanish, and that

$$\kappa_4 = -(n+1)/120.$$

45.3 In the $n!$ permutations of n numbers show that if P_n is the number with S positive differences,

$$P_n(S) = (S+1)P_{n-1}(S) + (n-S)P_{n-1}(S-1).$$

Hence obtain the recurrence expression for the moments

$$E_n(x^{2i+1}) = 0$$

$$E_n(x^{2i}) = \frac{n+1}{n}E_{n-1}(x+\frac{1}{2})^{2i} - \frac{2}{n}E_{n-1}(x+\frac{1}{2})^{2i+1}$$

where

$$x = S - E(S).$$

Hence by induction show that

$$\lim_{n \rightarrow \infty} \frac{\mu_{2i}(x)}{\{\mu_2(x)\}^i} = (2i-1)(2i-3) \dots 3.1$$

and thus that the distribution tends to normality.

(Mann, 1945a)

45.4 Show that, against the normal regression alternative, the turning-point test based on p defined at (45.2) has

$$\{E'(p)\}_{\beta=0} = 0, \quad \{E''(p)\}_{\beta=0} \neq 0,$$

and δ defined at (25.16) $= \frac{1}{4}$, so that its ARE, compared to the regression coefficient test, is zero; and that it is also zero compared to the difference-sign test.

(Stuart, 1954, 1956)

45.5 Observing that, in the notation of 45.23, a rank r_i may be expressed as

$$r_i = \sum_{j=1}^n H_{ij},$$

show that, for the normal regression alternative,

$$\left[\frac{\partial}{\partial \beta} E(r_i r_k) \right]_{\beta=0} = \frac{n(n+1)}{4\sqrt{n}} \{i+k-(n+1)\}.$$

Hence, for the statistic W of (45.50),

$$\left[\frac{\partial}{\partial \beta} E(W) \right]_{\beta=0} = 0, \quad \left[\frac{\partial^2}{\partial \beta^2} E(W) \right]_{\beta=0} \sim \frac{n^5}{24\pi},$$

and

$$\text{var } W \sim n^5/144,$$

and thus that W has $\delta_W = \frac{5}{4}$ and zero ARE compared to the regression estimator.

(Stuart, 1954, 1956)

45.6 If values $u_i, i = 1, 2, \dots, n$ are chosen at random from a set of continuous distributions with frequency functions $f_i(u)$, show that the expected number c of points of increase in the series u_1, u_2, \dots, u_n is given by

$$E(c) = \sum_{i=1}^{n-1} \int_{-\infty}^{\infty} f_{i+1}(u_{i+1}) \int_{-\infty}^{u_{i+1}} f_i(u_i) du_i du_{i+1}.$$

For the rectangular distribution with linear trend
 $f_i(u_i) = 1, \quad i\theta \leq u_i \leq i\theta + 1,$
 $= 0$ elsewhere,

show that

$$\lim_{n \rightarrow \infty} \frac{1}{n} E(c) = \frac{1}{2} \{1 + \theta(2 - |\theta|)\}, \quad -1 \leq \theta \leq 1,$$

$$= 0, \quad \theta < -1,$$

$$= 1, \quad \theta > 1.$$

(Levene, 1952)

45.7 As in Exercise 45.6, show that for the normal distribution

$$f_i(u_i) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}(u_i - \mu_i)^2 \right\}$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} E(c) = \Phi(\theta/\sqrt{2}),$$

where

$$\phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{1}{2}x^2} dx,$$

and the trend is given by $u_i = (i-1)\theta$.

(Levene, 1952, who tabulates the values)

45.8 For a random series x_1, x_2, \dots, x_n define

$$u_r = 1 \text{ if the } r\text{th observation is an upper record,}$$

$$= 0 \text{ otherwise;}$$

$$l_r = 1 \text{ if the } r\text{th observation is a lower record,}$$

$$= 0 \text{ otherwise;}$$

and

$$s_r = u_r + l_r, \quad d_r = u_r - l_r.$$

Define also

$$s = \sum_{r=2}^n s_r$$

$$d = \sum_{r=2}^n d_r.$$

The scoring commences at the second observation.

Then if $p^{(r)}(s, d)$ is the joint frequency function of s and d in a series of r observations, show that

$$p^{(r)}(s, d) = \left(1 - \frac{2}{r}\right) p^{(r-1)}(s, d) + \frac{1}{r} p^{(r-1)}(s-1, d-1) + \frac{1}{r} p^{(r-1)}(s-1, d+1)$$

$$p^{(1)}(0, 0) = 1$$

and hence derive the probability-generating function $g(\theta_1, \theta_2)$ as

$$g^{(n)}(\theta_1, \theta_2) = \frac{1}{n!} \prod_{r=0}^{n-2} (r + \theta_1 \theta_2 - \theta_1/\theta_2).$$

Hence derive, for the characteristic functions of s and d ,

$$\phi_s^{(n)}(t) = \frac{1}{n!} \prod_{r=0}^{n-2} (r + 2e^{it})$$

$$\phi_d^{(n)}(t) = \frac{1}{n!} \prod_{r=0}^{\infty} (r + 2 \cos t).$$

Derive also the joint c.f. and by inversion show that

$$p^{(n)}(s, d) = p_1^{(n)}(s) 2^{-s} \binom{s}{\frac{1}{2}(s+d)}$$

where p_1 is the frequency function of s given by

$$p_1^{(n)}(s) = \frac{1}{n!} 2^s u^{(n-2)}(n-s-1)$$

and $u^{(n)}(r)$ denotes the sum of products of all selections of r integers out of $1, 2, \dots, n$.
(Foster and Stuart, 1954)

45.9 In the foregoing exercise show that

$$E(d) = 0, \quad \text{var } d = 2 \sum_{r=2}^n \frac{1}{r},$$

$$E(s) = 2 \sum_{r=2}^n \frac{1}{r}, \quad \text{var } s = 2 \sum_{r=2}^n \frac{1}{r} - 4 \sum_{r=2}^n \frac{1}{r^2}.$$

(Foster and Stuart, 1954)

45.10 Show that r_{kc} of (45.55) cannot exceed unity, but that r_k of (45.53) may do so.

45.11 For the coefficient of (45.53) calculated from a random series show that

$$E(r_k) = -\frac{1}{n-1}$$

and hence that r_k is biased as an estimator of serial correlation in the parent series.

TIME-SERIES: TREND AND SEASONALITY

Determination of trend

46.1 It is an essential part of the concept of trend that the movement over fairly long periods is smooth. This means that we can represent the trend component, at least locally, by a polynomial in the time element t . Thus, given the series u_t , we may, in the first instance, seek for some polynomial

$$u_t = a_0 + a_1 t + a_2 t^2 + \dots + a_p t^p \quad (46.1)$$

which will give an account of the trend movement. By taking p great enough we can, of course, obtain as close a representation as we like to a finite series; and how large we take p is a matter for decision in particular cases.

We need not restrict ourselves to polynomials, although they are the most convenient mathematically. Any suitable function of the time can be taken, though we should naturally choose one which itself moved in a trend-like way. Growth curves, for example, may be represented by exponential functions, and population curves like that of Table 45.5 are sometimes represented by the logistic curve of type

$$u(t) = \frac{k}{1 + e^{-\lambda t}} \quad (46.2)$$

46.2 If a polynomial is fitted to the whole series by least squares, it evidently gives the curvilinear regression line of u_t on the variable t . It is, however, clear that to obtain a satisfactory trend-curve for data such as that of Table 45.4 (sheep population), we should have to take a polynomial of rather high order or a somewhat complicated more general function. This may appear somewhat artificial and in any case the coefficients of such a polynomial, being based on high-order moments, would be very unstable from the sampling viewpoint. A more practical objection, though by no means an unimportant one, is that if we add another term to the series, as for example if we are keeping an annual series up to date from year to year, the work of fitting has to be done afresh each time. Moreover, the trend-line may be affected throughout its length. When, therefore, the series has no very obvious trend it is more convenient to use the simpler methods described below.

Moving averages

46.3 An alternative to finding a polynomial which will represent the whole series is to determine a polynomial which will represent a part of it, and to use different polynomials for different parts. The simplest method, and one which forms the basis of the majority of methods of trend-fitting, is to take the first n terms (n being chosen at will), fit a polynomial of degree p , not greater than $n-1$, to them, and use that polynomial to determine the value in the middle of its range; then to repeat the operation

with the n terms from the second to the $(n+1)$ th, and so on, moving on one term at each stage. Unless other considerations require it, we take n to be odd, so that the middle point of the range corresponds in time to a value which is actually observed. Otherwise the middle point falls half-way between two observed values, or we have to use some value of the fitted polynomial other than the middle point, which results in a loss of useful symmetry.

46.4 Suppose, then, that the number of terms is chosen to be odd and is denoted by $2m+1$. Without loss of generality we may denote the terms by $u_{-m}, u_{-(m-1)}, \dots, u_0, \dots, u_{m-1}, u_m$. If we choose to fit to them a polynomial of degree p we may, in the usual way, determine the coefficients by least squares, i.e. solve the equations

$$\frac{\partial}{\partial a_j} \sum_{t=-m}^m (u_t - a_0 - a_1 t - \dots - a_p t^p)^2, \quad j = 0, 1, \dots, p, \quad (46.3)$$

which will give us equations typified by

$$\sum t^j u_t - a_0 \sum t^j - a_1 \sum t^{j+1} - \dots - a_p \sum t^{j+p} = 0. \quad (46.4)$$

The sums $\sum t^j$ are functions of m only. Thus, if we solve (46.4) for a_0 we shall find an equation of the form

$$a_0 = c_0 + c_1 u_{-m} + c_2 u_{-(m-1)} + \dots + c_{2m+1} u_m \quad (46.5)$$

where the c 's depend on m and p , but not on the u 's.

Now u_0 assumes the value a_0 at $t = 0$ and hence this value, as given by (46.5), is the value we require for the polynomial. As we see, this is equivalent to a weighted average of the observed values, the weights being independent of which part of the series is taken. Thus our process of fitting a trend-line consists of determining the constants c (which depend on m and p and therefore give us a twofold element of choice) and then calculating, for each consecutive set of $(2m+1)$ terms in the series, a value given by (46.5). If the terms are u_x, \dots, u_{2m+x} , the calculated value will correspond to $t = m+x$. A supplementary procedure is necessary to give values corresponding to the m terms at the beginning and the m terms at the end.

Example 46.1

Suppose we have a series and wish to fit a curve which best approximates to sets of seven points; and suppose we regard a cubic as providing a satisfactory approximation. What are the weights of the moving average?

We have $m = 3$ and $p = 3$, and our polynomial is

$$u_t = a_0 + a_1 t + a_2 t^2 + a_3 t^3.$$

Taking our origin at $t = 0$, we find, for equations (46.4), in virtue of the fact that $\sum t^k = 0$ for odd k ,

$$\left. \begin{aligned} \sum u &= 7a_0 && + 28a_2 \\ \sum tu &= &28a_1 && + 196a_3 \\ \sum t^2u &= 28a_0 && + 196a_2 \\ \sum t^3u &= &196a_1 && + 1588a_3 \end{aligned} \right\} \quad (46.6)$$

giving, for a_0 ,

$$a_0 = \frac{1}{2!} \{7 \sum u - \sum t^2 u\} \\ = \frac{1}{2!} \{-2u_{-3} + 3u_{-2} + 6u_{-1} + 7u_0 + 6u_1 + 3u_2 - 2u_3\}.$$

We may write this conveniently as

$$\frac{1}{2!} [-2, 3, 6, 7, 6, 3, -2]$$

or, when symmetrical formulae are used, as in the present case, by

$$\frac{1}{2!} [-2, 3, 6, 7],$$

denoting the middle term by heavy type.

To take a simple illustration. Suppose the series is given by the following values:

$t:$	1	2	3	4	5	6	7	8	9	10
$u_t:$	0	1	8	27	64	125	216	343	512	729

We have, for the trend-value at $t = 4$,

$$a_0 = \frac{1}{2!} \{(-2 \times 0) + (3 \times 1) + (6 \times 8) + \dots - (2 \times 216)\} \\ = \frac{1}{2!} .567 = 27.$$

The trend-value is equal to the actual value of the series, and this obviously must be so when we note that we are fitting a cubic to the series

$$u_t = (t-1)^3.$$

It will be observed that in this example we should have obtained the same value for a_0 if we fitted quadratics instead of cubics, for a_0 does not depend on a_3 in equations (46.6); and generally the case p odd includes the case of the next lowest (even) value of p , so that we need not give separate formulae for even p .

46.5 Writing $a_0[k]$ for the value of a_0 calculated in the above manner for an average of k successive terms, we find the following formulae up to $p = 5$. The reader may care to verify them for himself as an exercise. It will be evident that the sum of coefficients in any formula is unity; for if we apply the trend to a set of values all equal to unity, the result must be unity.

Quadratic and Cubic

$$\left. \begin{aligned} [5] & \quad \frac{1}{3!} [-3, 12, 17] \\ [7] & \quad \frac{1}{2!} [-2, 3, 6, 7] \\ [9] & \quad \frac{1}{2!} [-21, 14, 39, 54, 59] \\ [11] & \quad \frac{1}{4!} [-36, 9, 44, 69, 84, 89] \\ [13] & \quad \frac{1}{1!} [-11, 0, 9, 16, 21, 24, 25] \\ [15] & \quad \frac{1}{1!} [-78, -13, 42, 87, 122, 147, 162, 167] \\ [17] & \quad \frac{1}{3!} [-21, -6, 7, 18, 27, 34, 39, 42, 43] \\ [19] & \quad \frac{1}{2!} [-136, -51, 24, 89, 144, 189, 224, 249, 264, 269] \\ [21] & \quad \frac{1}{3!} [-171, -76, 9, 84, 149, 204, 249, 284, 309, 324, 329] \end{aligned} \right\} \quad (46.7)$$

Quartic and Quintic

$$\begin{aligned}
[7] & \frac{1}{2 \cdot 3 \cdot 1} [5, -30, 75, 131] \\
[9] & \frac{1}{4 \cdot 2 \cdot 9} [15, -55, 30, 135, 179] \\
[11] & \frac{1}{4 \cdot 2 \cdot 9} [18, -45, -10, 60, 120, 143] \\
[13] & \frac{1}{2 \cdot 4 \cdot 3 \cdot 1} [110, -198, -135, 110, 390, 600, 677] \\
[15] & \frac{1}{4 \cdot 6 \cdot 1 \cdot 8 \cdot 9} [2145, -2860, -2937, -165, 3755, 7500, 10125, 11063] \\
[17] & \frac{1}{4 \cdot 1 \cdot 9 \cdot 9} [195, -195, -260, -117, 135, 415, 660, 825, 883] \\
[19] & \frac{1}{7 \cdot 4 \cdot 2 \cdot 9} [340, -255, -420, -290, 18, 405, 790, 1110, 1320, 1393] \\
[21] & \frac{1}{2 \cdot 8 \cdot 0 \cdot 0 \cdot 1 \cdot 5} [11628, -6460, -13005, -11220, -3940, 6378, \\
& \quad 17655, 28190, 36660, 42120, 44003]
\end{aligned} \tag{46.8}$$

46.6 It is sometimes more convenient to express these formulae in terms of the differences of the series $\Delta^r u_t$ where

$$\Delta u_t = u_{t+1} - u_t. \tag{46.9}$$

Thus, for example,

$$\frac{1}{2 \cdot 1} [-2, 3, 6, 7, 6, 3, -2] = u_t - \frac{1}{2 \cdot 1} (9\Delta^4 + 9\Delta^5 + 2\Delta^6) u_{t-3} \tag{46.10}$$

which exhibits at once the fact that our fit is exact for a cubic, i.e. as far as fourth and higher differences. Or we may equally well represent the process as a moving average of the differences, which provides a convenient method of calculation when differences are smaller than lower-order differences or the original values of the series. For instance

$$\frac{1}{2 \cdot 1} [-2, 3, 6, 7] = u_t + \frac{1}{2 \cdot 1} \{2\Delta^3 u_{t-3} + 3\Delta^2 u_{t-2} - 3\Delta^3 u_{t-1} - 2\Delta^3 u_t\} \tag{46.11}$$

$$= u_t + \frac{1}{2 \cdot 1} [2, 3, -3, 2] \Delta^3 u_{t-3} \tag{46.12}$$

$$= u_t - \frac{1}{2 \cdot 1} [2, 5, 2] \Delta^4 u_{t-3}. \tag{46.13}$$

We can obviously represent such formulae in a variety of different ways. (46.13) is particularly convenient because it gives us the residuals immediately.

Example 46.2

Suppose we wish to represent one of the other formulae in (46.7) in this manner, say the quintic fitted to 11 points:

$$\frac{1}{4 \cdot 2 \cdot 9} [18, -45, -10, 60, 120, 143].$$

We first of all subtract unity from the middle term to give

$$\frac{1}{4 \cdot 2 \cdot 9} [18, -45, -10, 60, 120, -286]. \tag{46.14}$$

The sum of coefficients must now be zero. Denote by U a shift operator such that

$$Uu_t = u_{t+1}. \tag{46.15}$$

Then

$$\Delta = U - 1. \tag{46.16}$$

The moving average (46.14) may, apart from the divisor 429, be written

$$18 - 45U - 10U^2 + 60U^3 + 120U^4 - 286U^5 + 120U^6 + 60U^7 - 10U^8 - 45U^9 + 18U^{10}.$$

We know that this is exact as far as fifth differences and consequently $\Delta^6 = (U - 1)^6$ must be a factor. We find

$$(U - 1)^6 (18U^4 + 63U^3 + 98U^2 + 63U + 18).$$

The original process may then (since $U-1 = \Delta$) be written

$$u_t + \frac{1}{4 \cdot 2 \cdot 9} [18, 63, 98, 63, 18] \Delta^6 u_{t-5}. \quad (46.17)$$

46.7 The following are the formulae for (46.7) and (46.8) in terms of differences: (*)

$$\left. \begin{aligned} [5] \quad & u_3 - \frac{3}{3 \cdot 5} [1] \Delta^4 u_1 \\ [7] \quad & u_4 - \frac{1}{2 \cdot 1} [2, 5, 2] \Delta^4 u_2 \\ [9] \quad & u_5 - \frac{1}{2 \cdot 3 \cdot 1} [21, 70, 115, 70, 21] \Delta^4 u_3 \\ [11] \quad & u_6 - \frac{1}{4 \cdot 2 \cdot 9} [36, 135, 280, 385] \Delta^4 u_4 \\ [13] \quad & u_7 - \frac{1}{1 \cdot 4 \cdot 3} [11, 44, 101, 168, 210] \Delta^4 u_5 \\ [15] \quad & u_8 - \frac{1}{1 \cdot 1 \cdot 0 \cdot 5} [78, 325, 790, 1435, 2100, 2478] \Delta^4 u_6 \\ [17] \quad & u_9 - \frac{1}{3 \cdot 2 \cdot 3} [21, 90, 227, 434, 686, 924, 1050] \Delta^4 u_7 \\ [19] \quad & u_{10} - \frac{1}{2 \cdot 2 \cdot 6 \cdot 1} [136, 595, 1540, 3045, 5040, 7266, 9240, 10230] \Delta^4 u_8 \\ [21] \quad & u_{11} - \frac{1}{3 \cdot 0 \cdot 5 \cdot 9} [171, 760, 2005, 4060, 6930, 10416, 14070, 17160, \\ & \quad \quad \quad 18645] \Delta^4 u_9 \end{aligned} \right\} \quad (46.18)$$

Quartic and Quintic

$$\left. \begin{aligned} [7] \quad & u_4 + \frac{5}{2 \cdot 3 \cdot 1} [1] \Delta^6 u_1 \\ [9] \quad & u_5 + \frac{5}{4 \cdot 2 \cdot 9} [3, 7] \Delta^6 u_2 \\ [11] \quad & u_6 + \frac{1}{4 \cdot 2 \cdot 9} [18, 63, 98] \Delta^6 u_3 \\ [13] \quad & u_7 + \frac{1}{2 \cdot 4 \cdot 3 \cdot 1} [110, 462, 987, 1302] \Delta^6 u_4 \\ [15] \quad & u_8 + \frac{1}{4 \cdot 6 \cdot 1 \cdot 8 \cdot 9} [2145, 10010, 24948, 42273, 51198] \Delta^6 u_5 \\ [17] \quad & u_9 + \frac{1}{4 \cdot 1 \cdot 9 \cdot 9} [195, 975, 2665, 5148, 7623, 8778] \Delta^6 u_6 \\ [19] \quad & u_{10} + \frac{1}{7 \cdot 4 \cdot 2 \cdot 9} [340, 1785, 5190, 10875, 18018, 24453, 27258] \Delta^6 u_7 \\ [21] \quad & u_{11} + \frac{1}{2 \cdot 6 \cdot 0 \cdot 0 \cdot 1 \cdot 5} [11628, 63308, 192423, 426258, 759003, \\ & \quad \quad \quad 1135134, 1450449, 1581294] \Delta^6 u_8 \end{aligned} \right\} \quad (46.19)$$

46.8 Several methods have been proposed to simplify the arithmetic of fitting a trend-line by moving averages, the large numbers in some of the expressions in (46.7) and (46.8) involving considerable labour in straightforward application. The simplest, perhaps, is that of iterated averages.

Suppose we take an average of sets of four with equal weights—a very simple process—and then another average of the same kind of that average. If the primary series is u_t , the result of the first operation will be to give a series typified by

$$v_1 = \frac{1}{4}(u_1 + u_2 + u_3 + u_4)$$

and that of the second operation to give

$$w_1 = \frac{1}{4}(v_1 + v_2 + v_3 + v_4)$$

$$= \frac{1}{16}(u_1 + 2u_2 + 3u_3 + 4u_4 + 3u_5 + 2u_6 + u_7).$$

We may write this symbolically as

$$\left\{ \frac{1}{4} [1, 1, 1, 1] \right\}^2 = \frac{1}{16} [1, 2, 3, 4] \quad (46.20)$$

$$\left\{ \frac{1}{4} [1, 1, 1, 1] \right\}^2 = \frac{1}{16} [1, 2, 3, 4] \quad (46.21)$$

(*) Kendall (1961a) has shown that the numbers in square brackets on the right in these formulae tend, as n and p increase, to the ordinates of a normal frequency function.

or, reserving the symbol $\frac{1}{k}[k]$ for a simple arithmetic mean of terms, as

$$\frac{1}{16}[4]^2 = \frac{1}{16}[1, 2, 3, 4]. \quad (46.22)$$

Now compare the weights of the average derived in Example 46.1 for fitting a cubic to seven points. Reduced to unit divisors, we have for the weights of the latter

$$-0.0952, 0.1429, 0.2857, 0.3333$$

and for the weights of (46.20)

$$0.0625, 0.1250, 0.1875, 0.2500.$$

The two are not identical, but they follow the same sort of course and it might be possible to regard the latter as an approximation to the former. (We shall derive better approximations presently, but this will serve for purposes of illustration.) Now the iterated summation resulting in (46.20) is much easier to carry out than the single weighted averaging process of Example 46.1. Generally, if we can find averages with simple integral weights, preferably unity, which will, in conjunction, give approximations to the more complicated weights of a single average, it is usually easier to use the iteration process.

46.9 In the notation of finite differences, write

$$\delta u_t = u_{t+\frac{1}{2}} - u_{t-\frac{1}{2}}. \quad (46.23)$$

We have, for the second "central" difference $\delta^2 u_t$,

$$\begin{aligned} \delta^2 u_t &= (u_{t+1} - u_t) - (u_t - u_{t-1}) \\ &= (U - 2 + U^{-1})u_t. \end{aligned} \quad (46.24)$$

Writing

$$U = \exp(2i\phi), \quad (46.25)$$

we find, symbolically,

$$\begin{aligned} \delta^2 &= \exp(2i\phi) - 2 + \exp(-2i\phi) \\ &= -4 \sin^2 \phi. \end{aligned} \quad (46.26)$$

Then

$$\begin{aligned} \sum_{j=-m}^m u_j &= \sum_{j=-m}^m U^j u_0 \\ &= \left(1 + 2 \sum_{j=1}^m \cos 2j\phi\right) u_0, \end{aligned}$$

since the terms in $\sin 2j\phi$ vanish,

$$= \frac{\sin(2m+1)\phi}{\sin \phi} u_0. \quad (46.27)$$

Thus

$$\begin{aligned} \frac{1}{k}[k]u_0 &= \frac{1}{k} \frac{\sin k\phi}{\sin \phi} u_0 \\ &= \frac{1}{k} \left\{ k - \frac{k(k^2-1)}{3!} \sin^2 \phi + \frac{k(k^2-1)(k^2-3^2)}{5!} \sin^4 \phi - \dots \right\} u_0 \\ &= u_0 + \frac{k^2-1}{2^2 3!} \delta^2 u_0 + \frac{(k^2-1)(k^2-3^2)}{2^4 5!} \delta^4 u_0 + \dots \end{aligned} \quad (46.28)$$

This interesting formula gives the arithmetic average in terms of the middle term u_0 and its central differences.

If now our series is approximately represented by a cubic, so that fourth differences vanish, we have, taking u_0 as the middle term,

$$\frac{1}{k}[k]u_0 = u_0 + \frac{k^2-1}{24}\delta^2 u_0, \quad (46.29)$$

and this equation will in any case be true up to third differences. Similarly, for two iterated averages we have, to the same order,

$$\frac{1}{k_1 k_2}[k_1][k_2]u_0 = u_0 + \frac{1}{24}(k_1^2 + k_2^2 - 2)\delta^2 u_0, \quad (46.30)$$

and so on. We will use these results to derive two formulae in very general use by actuaries for "graduating" a series, a process which is very similar to that of fitting a trend-line.

Example 46.3 Spencer's 15-point formula

Consider three successive averages with equal weights

$$\begin{aligned} \frac{1}{80}[4][4][5]u_0 &= u_0 + \frac{1}{24}(4^2 + 4^2 + 5^2 - 3)\delta^2 u_0 \\ &= u_0 + \frac{9}{4}\delta^2 u_0. \end{aligned}$$

Multiplying by $1 - \frac{9}{4}\delta^2$, we then have, to third differences,

$$u_0 = \frac{1}{80}[4]^2[5](1 - \frac{9}{4}\delta^2)u_0.$$

Substituting for δ^2 the formula $[1, -2, 1]$, as given by (46.24), we find

$$u_0 = \frac{1}{320}[4]^2[5][-9, 22, -9].$$

Now without affecting the order of the approximation we may add factors in δ^4 or higher central differences, and can simplify the numerical coefficients to some extent. Let us add to the factor $[-9, 22, -9]$ a term $-3\delta^4 = [-3, 12, -18, 12, -3]$. The result is $[-3, 3, 4, 3, -3]$, giving

$$u_0 = \frac{1}{320}[4]^2[5][-3, 3, 4, 3, -3]. \quad (46.31)$$

This is Spencer's 15-point formula. It covers sets of 15 consecutive terms, the weights in full being

$$\frac{1}{320}[-3, -6, -5, 3, 21, 46, 67, 74]. \quad (46.32)$$

Example 46.4 Spencer's 21-point formula

In a similar way we find

$$\frac{1}{175}[5]^2[7] = 1 + 4\delta^2$$

giving, to third differences,

$$u_0 = \frac{1}{175}[5]^2[7][-4, 9, -4]u_0.$$

We now add to the factor $[-4, 9, -4]$ the expression

$$-3\delta^4 - \frac{1}{2}\delta^6 = [-3, 12, -18, 12, -3] + [-\frac{1}{2}, 3, -7\frac{1}{2}, 10, 7\frac{1}{2}, 3, -\frac{1}{2}]$$

giving

$$\begin{aligned} u_0 &= \frac{1}{175}[5]^2[7][-\frac{1}{2}, 0, \frac{1}{2}, 1] \\ &= \frac{1}{350}[5]^2[7][-1, 0, 1, 2]. \end{aligned} \quad (46.33)$$

This is Spencer's 21-point formula.

46.10 Simplicity of calculation, however, is not nowadays as important as it once was, and for certain purposes these approximations are to be avoided. The original formulae (46.7) and (46.8) provide lines of closest fit for assigned extent and degree of polynomial. It follows that for given m and p the sum of squares of weighting coefficients is a minimum. In fact, if we apply the moving average to a series consisting of a polynomial trend of degree p plus a random residual ε_t (which has the same distribution for all t) the residual sum of squares is given by

$$\Sigma (c_0 \varepsilon_1 + c_1 \varepsilon_{j+1} + \dots + c_{2m} \varepsilon_{j+2m})^2,$$

so that the expected variance of residuals is

$$\sum_{j=0}^{2m} c_j^2 \quad (46.34)$$

which is thus a minimum within the class of weights reproducing the same degree of polynomial for assigned m .

End-effects

46.11 The moving-average method as we have expounded it has obvious properties of symmetry. It also has the drawback of failing to provide trend-values for the first m and the last m terms of the series. As a rule it is not a great loss to have to forgo the values at the beginning, but the absence of trend-values at the end is a serious handicap, especially when we want to extrapolate into the future. We can fill the gap in various ways, recognizing that trend-values at the end may not be so reliable as those in the middle. The method illustrated in the following example is probably as simple as any.

Example 46.5

Consider again the formula used in Example 46.1,

$$\frac{1}{21}[-2, 3, 6, 7].$$

We obtained this by fitting a cubic, but used that cubic only to determine the middle point of a set of seven. There is no reason why we should not also use it to determine the last three points of the end-set of seven. To do so, however, we need to solve equations (46.6) for a_1, a_2, a_3 as well as for a_0 .

We find

$$\begin{aligned} a_1 &= \frac{1}{1512} \{397 \Sigma tu - 49 \Sigma t^3 u\}, \\ a_2 &= \frac{1}{84} \{-4 \Sigma u + \Sigma t^2 u\}, \\ a_3 &= \frac{1}{216} \{-7 \Sigma tu + \Sigma t^3 u\}. \end{aligned}$$

From these results substituted in the polynomial we find, in an obvious notation,

$$\begin{aligned} u_t &= \frac{1}{21}[-2, 3, 6, 7, 6, 3, -2] + \frac{t}{252}[22, -67, -58, 0, 58, 67, -22] \\ &\quad + \frac{t^2}{84}[5, 0, -3, -4, -3, 0, 5] + \frac{t^3}{36}[-1, 1, 1, 0, -1, -1, 1]. \end{aligned} \quad (46.35)$$

Thus, for example, with $t = 1, 2, 3$, the expressions reduce to

$$u_1 = \frac{1}{42}[1, -4, 2, 12, 19, 16, -4], \quad (46.36)$$

THE ADVANCED THEORY OF STATISTICS

$$u_2 = \frac{1}{42}[4, -7, -4, 6, 16, 19, 8], \quad (46.37)$$

$$u_3 = \frac{1}{42}[-2, 4, 1, -4, -4, 8, 39]. \quad (46.38)$$

For example, if our last seven terms were 0, 1, 8, 27, 64, 125, 216, the fourth trend term (the one following the middle term, where $t = 0$) would be

$$\frac{1}{42}[0 - 4 + 16 + 324 + 1216 + 2000 - 864] = 64.$$

The next is

$$\frac{1}{42}[0 - 7 - 32 + 162 + 1024 + 2375 + 1728] = 125$$

and the reader can verify that the last is 216. These results, of course, are exact because we are fitting a cubic to a cubic.

One interesting phenomenon is to be noted here. In (46.35) to (46.38), as in the formula for u_0 , the coefficients sum to unity. But they become more and more unequal so that their sum of squares, in general, increases as we go from u_0 to u_3 . The variance of any random residual term therefore increases, a reflection of the fact that, as we depart more from the centre of the range which we have fitted, the polynomials become less "reliable." The sums of squares of coefficients in this case are

$$u_0: 0.3333, \quad u_1: 0.4524, \quad u_2: 0.4524, \quad u_3: 0.9286.$$

The increase in sum of squares is not, however, monotonic. Some further results are given in Exercises 46.1 and 46.15.

The coefficients and their sums of squares have been tabulated by Cowden (1962) for $p \leq 5$, $n \leq 25$.

46.12 Results of the foregoing kind may also very conveniently be obtained by the use of orthogonal polynomials (28.18, Vol. 2). If we put $n = 2m + 1$ we find, for the first four polynomials,

$$\left. \begin{aligned} \phi_0(t) &= 1 \\ \phi_1(t) &= \lambda_1 t \\ \phi_2(t) &= \lambda_2 \{t^2 - \frac{1}{3}m(m+1)\} \\ \phi_3(t) &= \lambda_3 \{t^3 - \frac{1}{5}t(3m^2 + 3m - 1)\} \\ \phi_4(t) &= \lambda_4 \{t^4 - \frac{1}{7}t^2(6m^2 + 6m - 5) + \frac{3}{35}(m-1)(m)(m+1)(m+2)\} \end{aligned} \right\} \quad (46.39)$$

If now, for example,

$$u_t = b_0 + b_1 \phi_1 + b_2 \phi_2 + b_3 \phi_3 \quad (46.40)$$

we have at once

$$u_0 = b_0 - \frac{1}{3}\lambda_2 m(m+1)b_2. \quad (46.41)$$

Moreover, from the tables of the function the values of λ can be obtained and also the values $\Sigma \phi_i^2$. We have, in virtue of the orthogonality property,

$$b_0 = \frac{\Sigma u_t \phi_0}{\Sigma \phi_0^2} = \frac{\Sigma u_t}{2m+1} \quad (46.42)$$

$$b_2 = \frac{\Sigma u_t \phi_2}{\Sigma \phi_2^2}. \quad (46.43)$$

For example, with a cubic fitted to sets of 7, $m = 3$, and we have, from (46.42), (46.43)

and the tables,

$$b_0 = \frac{1}{7} \sum u_t, \quad b_2 = \frac{1}{84} \sum (t^2 - 4) u_t.$$

Hence, from (46.40),

$$\begin{aligned} u_0 &= \frac{1}{7} \sum u_t - \frac{1}{21} (\sum t^2 u_t - 4 \sum u_t) \\ &= \frac{1}{21} (7 \sum u_t - \sum t^2 u_t), \end{aligned}$$

as in Example 46.1. A similar use of the polynomials will give us the end-values discussed in 46.11.

46.13 We have as yet said nothing about criteria by which we should decide the extent of a moving average, $2m+1$, or the degree of the polynomial, p , on which it should be based. There are, in fact, no simple criteria of this kind. One important reason for this is that a great deal depends on why we are interested in isolating the trend, or, to put the matter in a rather different way, what is the underlying model which determines our dissection of the series. If we are concerned chiefly to describe a broad trend in the data, and are not particularly interested in short-term and residual effects, one type of moving average may be adequate. But if we want to remove the trend in order to study the residuals, such a type may be quite inappropriate; and indeed, for some purposes, we may well question whether it is safe to eliminate trend by a moving average at all. Before, then, we can adequately discuss the choice of a suitable method of finding the trend we must consider the effect of our methods on residual variation.

The effect of trend-elimination by moving averages on other components

46.14 In Table 46.1 we have applied the Spencer 21-point formula to an artificial series obtained by adding a random element to a cubic. (We have chosen this formula rather than one of (46.7) because the effect of successive simple averages can also be seen.) Specifically,

$$u_t = (t-26) + \frac{1}{10}(t-26)^2 + \frac{1}{100}(t-26)^3 + \varepsilon_t. \quad (46.44)$$

The component ε_t was taken from tables of random numbers and consists of samples from a population in which all integral values from 0 to 99 are equally frequent. The various columns of the table illustrate the process of fitting, and we may note in passing that for a series as short as this it is convenient to leave the more difficult summations to the last as there are substantially fewer of them.

Now we know that the Spencer formula will fit a cubic exactly, so that when we subtract the trend from the original series we ought to eliminate the systematic constituent entirely and be left with our random component, except in so far as we have rounded off the systematic element to the nearest unit. A comparison of columns (2) and (9) in Table 46.1, remembering that the latter includes an element 49.5 equal to the mean of the random component, shows that we do not do so. The reason is not far to seek. The moving average has acted on the random element itself and determined a "trend-line" in it.

The results of applying the Spencer 21-point formula to the random element ε_t are shown in column (11). We should expect that if the method were perfect the values in this column would be 49.5, the mean of ε_t , apart from irregular sampling effects;

THE ADVANCED THEORY OF STATISTICS

376

Table 46.1—Series given by equation (46.44) with trend-line determined by a Spencer 21-point formula

(1) t	(2) cubic term	(3) ϵ_t	(4) u_t	(5) [5] u_t	(6) [5] (5)	(7) [7] (6)	(8) [-1, 0, 1, 2, ...] (7)	(9) $\frac{1}{250}$ (8)	(10) Deviation $u_t - (9)$	(11) Graduation of ϵ_t alone
1	-119	23	-96							
2	-105	15	-90							
3	-92	75	-17	-246						
4	-80	48	-32	-209						
5	-70	59	-11	-87	-572					
6	-60	1	-59	-42	-241					
7	-51	83	32	12	162					
8	-44	72	28	85	413	2233				
9	-37	59	22	194	670	3801				
10	-31	93	62	164	844	5120				
11	-26	76	50	215	957	5984	14352	41	9	67
12	-22	24	2	186	996	6642	15470	44	-42	66
13	-18	97	79	198	1078	7041	15815	45	34	63
14	-15	8	-7	233	1026	7145	15676	45	-52	60
15	-12	86	74	246	1071	7038	14978	43	31	55
16	-10	95	85	163	1069	6934	14166	40	45	51
17	-8	23	15	231	984	6709	13379	38	-23	47
18	-7	3	-4	196	850	6535	12703	36	-40	43
19	-6	67	61	112	892	6408	12169	35	26	40
20	-5	44	39	148	853	6363	12102	35	4	39
21	-4	5	1	205	852	6446	12279	35	-34	39
22	-3	54	51	192	944	6611	12676	36	15	39
23	-2	55	53	195	1024	6769	13228	38	15	40
24	-2	50	48	204	1031	7052	13857	40	8	41
25	-1	43	42	228	1015	7353	14508	41	1	42
26	0	10	10	212	1050	7610	15120	43	-33	43
27	1	74	75	176	1136	7923	15634	45	30	44
28	2	35	37	230	1153	8249	16251	46	-9	44
29	4	8	12	290	1201	8607	17002	49	-37	45
30	6	90	96	245	1337	9019	17717	51	45	44
31	9	61	70	260	1357	9424	18499	53	17	44
32	12	18	30	312	1373	9870	19307	55	-25	43
33	15	37	52	250	1462	10429	20159	58	-6	42
34	20	44	64	306	1541	10989	21133	60	4	41
35	24	10	34	334	1599	11679	22417	64	-30	39
36	30	96	126	339	1760	12539	23797	68	58	38
37	36	22	58	370	1897	13529	25737	74	-16	37
38	44	13	57	411	2047	14699	27955	80	-23	36
39	52	43	95	443	2233	16060	30456	87	8	35
40	61	14	75	484	2452	17570	33334	95	-20	34
41	71	87	158	525	2711	19353	36716	105	53	34
42	83	16	99	589	2960	21394				
43	95	3	98	670	3270	23690				
44	109	50	159	692	3680	26255				
45	124	32	156	794	4088					
46	140	40	180	935	4529					
47	158	43	201	997	5017					
48	177	62	239	1111						
49	198	23	221	1180						
50	240	50	270							
51	244	5	249							

but not only do the observed values deviate from this mean, they do so systematically, the values having a small oscillatory movement which is shown as part of the trend.

46.15 This effect is vital, particularly if we are eliminating trend so as to concentrate attention on oscillations. We proceed to examine it more closely.

Suppose that we have a series composed of the sum of three parts, a trend $x_1(t)$, an oscillatory term $x_2(t)$, and a random element $x_3(t)$, so that

$$u_t = x_1 + x_2 + x_3. \quad (46.45)$$

If we determine the trend by a moving average, denoted by an operation T , then clearly

$$Tu_1 = Tx_1 + Tx_2 + Tx_3. \quad (46.46)$$

Let us now suppose that our method of determining trend is perfect in the sense that $Tx_1 = x_1$. Then, on subtracting (46.46) from (46.45) to eliminate trend, we find

$$u_t - Tu_t = x_2 - Tx_2 + x_3 - Tx_3. \quad (46.47)$$

The point of present interest is that the terms Tx_2 and Tx_3 in (46.47) may distort the genuinely oscillatory parts of the residual series and induce spurious oscillatory movements.

46.16 Consider the simple case when x_2 is a sine term, $\sin(\alpha + \lambda t)$, t being integral. Since

$$\sum_{t=1}^k \sin(\alpha + \lambda t) = \frac{\sin \frac{1}{2}k\lambda}{\sin \frac{1}{2}\lambda} \sin \left\{ \alpha + \frac{1}{2}(k+1)\lambda \right\}, \quad (46.48)$$

a simple moving average of k consecutive terms will result in a sine series of the same period and phase as the original, but with the amplitude reduced by the factor

$$\frac{1}{k} \frac{\sin \frac{1}{2}k\lambda}{\sin \frac{1}{2}\lambda}. \quad (46.49)$$

Iteration q times will reduce the amplitude by the q th power of this factor.

Thus the term Tx_2 will be small if k is large, q is large, or if $\frac{1}{2}k\lambda$ is a multiple of π , that is, if the extent of the moving average is a period of the oscillation. But if λ is small and $k\lambda$ is small the amplitude is reduced very little and $x_2 - Tx_2$ will largely disappear, i.e. the moving average will partially obliterate the term in x_2 . In this case, $k\lambda$ being small, the extent of the moving average is small compared with the period of the harmonic term, that is to say the oscillation is a slow one. This result is what we should expect. A slow oscillation is treated as a trend by the moving average and eliminated accordingly. Generally, the moving average will emphasize the shorter oscillations at the expense of the longer ones. Furthermore, if the extent of the average is slightly greater than the period, the term (46.49) may have a negative sign, and consequently the difference from the trend may somewhat exaggerate the true oscillations.

It is not so easy to exhibit the precise effect of the moving average when the weights are unequal and the terms are not harmonic, but evidently the same kind of situation is apt to arise.

46.17 Now consider the effect of a simple moving average (that is, one with equal weights) on the residual element x_3 , which we will suppose to be a random element ε_t with variance v . For the term Tx_3 we have

$$Tx_3 = \frac{1}{k} \sum_{j=-[\frac{1}{2}k]}^{[\frac{1}{2}k]} \varepsilon_{t+j} \quad (46.50)$$

where $[\frac{1}{2}k]$ is the greatest integer which does not exceed $\frac{1}{2}k$. Consecutive values of ε_t are independent, but consecutive values of Tx_3 are not; for $Tx_3(a)$ and $Tx_3(b)$ have $k-(a-b)$ values of ε in common and are correlated if $a-b < k$. Thus the series Tu_3 will be much smoother than u_3 , and if we proceed to further averagings will become smoother still. We have had an example of this effect in Table 46.1 and shall meet further examples below.

46.18 The effect of taking a moving average of a random series will then be to generate an oscillatory series, provided that the weights are such as to give a positive correlation between successive members of the generated series, a condition which is always realized in moving averages employed for trend-fitting. We shall call this the Slutsky-Yule effect, after the two statisticians who (independently) studied it in detail.

The generated series is not regular in the cyclical sense, that is to say its peaks and troughs do not recur at equal intervals of time, and the amplitudes of the oscillations vary considerably (although, in Chapter 49, we shall prove a theorem of Slutsky's showing that certain kinds of iterated average generate a sine curve). Nevertheless such oscillations present a striking resemblance to the kind of movement which is found in practice, particularly in economic time-series, and we shall consider them in more detail later. For our present purposes we require to consider how far the process of trend-elimination itself may generate such effects, in order to be sure that oscillatory movements in a trend-free series have not been put there, so to speak, by our own arithmetical processes.

46.19 For this purpose we shall consider the period and variance of a series generated by the Slutsky-Yule effect.

Since the peaks and troughs do not recur at equal intervals there is no quantity which we can conveniently call the length of the oscillation. There will, in fact, be a distribution of lengths. We may define as the mean length either the mean period from peak to peak, or that from trough to trough; but this raises some difficulties as to whether we are prepared to admit as periods small ripples on the main undulation.

Recognizing its somewhat arbitrary character, we shall take as our measure of oscillatory length the mean distance between "upcrosses," that is to say the mean distance between points where the series changes sign from negative to positive or "crosses the u -axis." Suppose the series is generated by a moving average with weights a_1, \dots, a_k of a random variable which is normally distributed with variance v . Then the probability that

$$u_k = \sum_{j=1}^k a_j \varepsilon_j < 0 \quad (46.51)$$

and

$$u_{k+1} = \sum_{j=1}^k a_j \varepsilon_{j+1} > 0, \quad (46.52)$$

i.e. that the generated series changes sign from negative to positive, is the proportional frequency of

$$dF = \frac{1}{(2\pi)^{\frac{1}{2}(k+1)}} \exp \left\{ -\frac{1}{2v} \sum_{j=1}^{k+1} \varepsilon_j^2 \right\} d\varepsilon_1 \dots d\varepsilon_{k+1} \quad (46.53)$$

between the hyperplanes $\sum a_j \varepsilon_j = 0$ and $\sum a_j \varepsilon_{j+1} = 0$. This is equal to the angle between these two planes, which is given by

$$\cos \theta = \frac{\sum_{j=1}^k a_j a_{j+1}}{\sum_{j=1}^k a_j^2}. \quad (46.54)$$

Hence the mean distance between upcrosses is $2\pi/\theta$, where θ is given by (46.54).

46.20 In a similar way, the probability that

$$u_{k+1} - u_k < 0 \quad (46.55)$$

$$u_k - u_{k-1} > 0, \quad (46.56)$$

that is, that u_k is a peak of the series, is the angle between the two hyperplanes

$$\sum_{j=1}^k a_j \varepsilon_{j+1} - \sum_{j=1}^k a_j \varepsilon_j = 0 \quad (46.57)$$

$$\sum_{j=1}^k a_j \varepsilon_j - \sum_{j=1}^k a_j \varepsilon_{j-1} = 0 \quad (46.58)$$

and is given by

$$\cos \theta_1 = \frac{(a_2 - a_1)a_1 + (a_3 - a_2)(a_2 - a_1) + \dots + a_k(a_k - a_{k-1})}{a_1^2 + (a_2 - a_1)^2 + \dots + a_k^2}. \quad (46.59)$$

Thus the mean distance between peaks is $2\pi/\theta_1$. The same formula obviously applies to mean distance between troughs.

46.21 If we wish to exclude "ripples" of a certain length d from consideration, we may enquire for the probability that (46.57) and (46.58) are satisfied in conjunction with

$$u_k > u_{k+d}. \quad (46.60)$$

This is evidently the area cut off on the unit sphere by the three planes (46.57), (46.58) and

$$\sum a_j \varepsilon_j - \sum a_j \varepsilon_{j+d} = 0. \quad (46.61)$$

If the angles between the planes are A , B and C , this area is $A + B + C - 2\pi = \theta_2$, say. The mean length between peaks, ripples excepted, is then $4\pi/\theta_2$.

Example 46.6

In Table 46.2 we show 480 terms of a series of random numbers which can take integral values from 0 to 19, together with a moving sum of fives of a moving sum of threes. Fig. 46.1 shows a portion of the derived series graphically. There are 474 terms of the smoothed series.

Table 46.2—Series of 480 terms of a rectangular random series ϵ and a [5] [3] smoothing S

Table 46.2—Series of 480 terms																				
[5] [3] smoothing																				
t	ϵ	S	t	ϵ	S	t	ϵ	S	t	ϵ	S	t	ϵ	S	t	ϵ	S	t	ϵ	S
1	3		81	17	197	161	9	140	241	1	99	321	17	196	401	17	191			
2	15		82	11	200	162	14	122	242	17	80	322	15	209	402	3	205			
3	15		83	6	206	163	1	117	243	0	75	323	13	194	403	18	198			
4	8	164	84	17	215	164	11	94	244	3	73	324	10	179	404	14	192			
5	19	147	85	18	228	165	1	98	245	0	94	325	18	151	405	14	191			
6	1	143	86	19	230	166	8	93	246	6	124	326	0	133	406	13	197			
7	3	145	87	15	220	167	2	106	247	17	169	327	9	112	407	5	205			
8	12	165	88	13	198	168	18	103	248	16	195	328	8	108	408	19	204			
9	19	175	89	8	175	169	1	121	249	17	204	329	3	105	409	17	202			
10	13	196	90	10	159	170	7	117	250	15	191	330	9	111	410	18	192			
11	16	191	91	14	158	171	9	127	251	3	175	331	12	107	411	5	174			
12	4	178	92	5	158	172	13	120	252	14	150	332	3	101	412	7	140			
13	17	159	93	12	159	173	2	137	253	9	144	333	8	85	413	15	107			
14	8	150	94	18	153	174	16	139	254	11	131	334	5	77	414	1	86			
15	6	134	95	1	145	175	1	145	255	3	135	335	1	75	415	4	66			
16	15	118	96	14	124	176	17	142	256	15	125	336	2	93	416	2	58			
17	3	101	97	8	112	177	13	145	257	1	138	337	13	107	417	3	50			
18	3	88	98	1	108	178	0	149	258	14	142	338	6	134	418	2	62			
19	7	87	99	5	123	179	15	149	259	9	162	339	15	151	419	10	78			
20	4	100	100	13	131	180	7	166	260	13	166	340	7	161	420	0	105			
21	5	126	101	11	150	181	16	167	261	11	182	341	13	160	421	9	126			
22	14	140	102	14	151	182	16	171	262	15	190	342	13	162	422	16	146			
23	15	147	103	6	140	183	7	169	263	8	203	343	5	155	423	9	152			
24	10	150	104	13	120	184	6	174	264	17	210	344	15	153	424	11	144			
25	3	153	105	1	119	185	13	168	265	19	214	345	10	162	425	12	124			
26	10	156	106	4	120	186	17	170	266	10	211	346	3	174	426	2	106			
27	13	165	107	13	133	187	14	170	267	17	188	347	18	176	427	3	106			
28	14	175	108	13	147	188	2	159	268	11	163	348	19	177	428	9	119			
29	15	168	109	8	172	189	15	140	269	9	146	349	8	174	429	6	139			
30	8	160	110	12	186	190	9	139	270	1	153	350	5	173	430	17	159			
31	10	154	111	12	195	191	1	145	271	11	154	351	16	159	431	15	174			
32	1	156	112	19	204	192	15	151	272	17	162	352	7	157	432	5	179			
33	18	154	113	13	203	193	16	147	273	17	155	353	16	157	433	14	172			
34	17	165	114	11	184	194	6	144	274	4	154	354	8	169	434	14	155			
35	4	164	115	18	156	195	10	132	275	8	137	355	6	168	435	9	133			
36	10	159	116	2	135	196	5	128	276	2	134	356	15	165	436	8	107			
37	16	138	117	4	121	197	7	122	277	18	141	357	19	153	437	3	75			
38	2	137	118	10	111	198	8	126	278	8	172	358	4	150	438	1	53			
39	13	131	119	8	116	199	18	120	279	9	184	359	5	133	439	3	55			
40	3	140	120	10	131	200	0	121	280	19	185	360	9	120	440	1	72			
41	14	135	121	3	145	201	7	105	281	17	167	361	12	117	441	5	91			
42	7	146	122	16	156	202	9	99	282	4	150	362	2	127	442	16	96			
43	16	141	123	12	173	203	5	93	283	8	123	363	11	118	443	8	91			
44	3	139	124	8	175	204	3	95	284	5	115	364	12	112	444	2	78			
45	10	117	125	19	160	205	12	91	285	6	131	365	5	105	445	0	75			
46	12	96	126	11	145	206	4	93	286	7	168	366	0	100	446	2	85			
47	0	75	127	1	129	207	2	97	287	19	186	367	12	84	447	7	109			
48	3	65	128	4	123	208	11	97	288	19	196	368	6	88	448	17	124			
49	2	61	129	16	108	209	6	107	289	16	188	369	2	96	449	12	124			
50	3	71	130	3	115	210	7	115	290	2	180	370	4	104	450	5	117			
51	10	84	131	13	108	211	6	128	291	10	158	371	15	109	451	2	106			
52	5	91	132	0	118	212	15	125	292	12	147	372	6	130	452	2	97			
53	10	92	133	10	112	213	4	130	293	15	145	373	5	148	453	15	92			
54	3	101	134	4	122	214	13	126	294	5	148	374	14	156	454	8	100			
55	2	119	135	19	113	215	4	125	295	6	145	375	14	164	455	2	111			
56	11	141	136	3	110	216	7	123	296	15	134	376	11	180	456	4	120			
57	14	166	137	4	100	217	13	119	297	6	137	377	8	187	457	11	121			
58	18	190	138	7	103	218	4	111	298	13	136	378	15	174	458	15	119			
59	8	212	139	0	106	219	13	101	299	2	136	379	18	151	459	8	110			
60	14	211	140	16	107	220	0	91	300	14	129	380	7	127	460	3	98			
61	15	204	141	13	102	221	3	82	301	9	128	381	1	99	461	1	98			
62	17	191	142	0	103	222	11	76	302	4	115	382	2	88	462	4	121			
63	7	185	143	2	114	223	0	75	303	14	100	383	7	89	463	13	150			
64	9	166	144	4	127	224	10	72	304	2	93	384	4	119	464	17	170			
65	11	160	145	18	136	225	1	86	305	0	96	385	18	143	465	19	176			
66	14	167	146	18	140	226	4	92	306	8	109	386	5	166	466	5	169			
67	5	188	147	6	131	227	6	109	307	12	131	387	17	170	467	4	149			
68	17	203	148	0	121	228	18	116	308	10	159	388	12	179	468	15	136			
69	18	204	149	4	120	229	3	139	309	11	178	389	7	179	469	8	137			
70	13	205	150	11	137	230	7	149	310	17	187	390	14	184	470	6	136			
71	18	185	151	15	162	231	12	149	311	10	200	391	15	190	471	14	133			
72	0	171	152	15	179	232	15	141	312	12	212	392	11	194	472	9				

The mean value of our series is $15 \times 9.5 = 142.5$. The number of upcrosses will be found from the table to be 23, the first between the 19th and 20th term of the smoothed series, the last between the 459th and the 460th. The mean distance between upcrosses is then $440/22 = 20$ units. How does this compare with the mean distance given by "normal" theory?

The weights of the graduation are $[1, 2, 3, 3, 3, 2, 1]$ and from (46.54) we have

$$\cos \theta = \frac{3.4}{3.7} = 0.9189$$

$$\theta = 23^\circ 14'.$$

Hence the mean distance = $360/23.233 = 15.5$ units. The observed mean distance

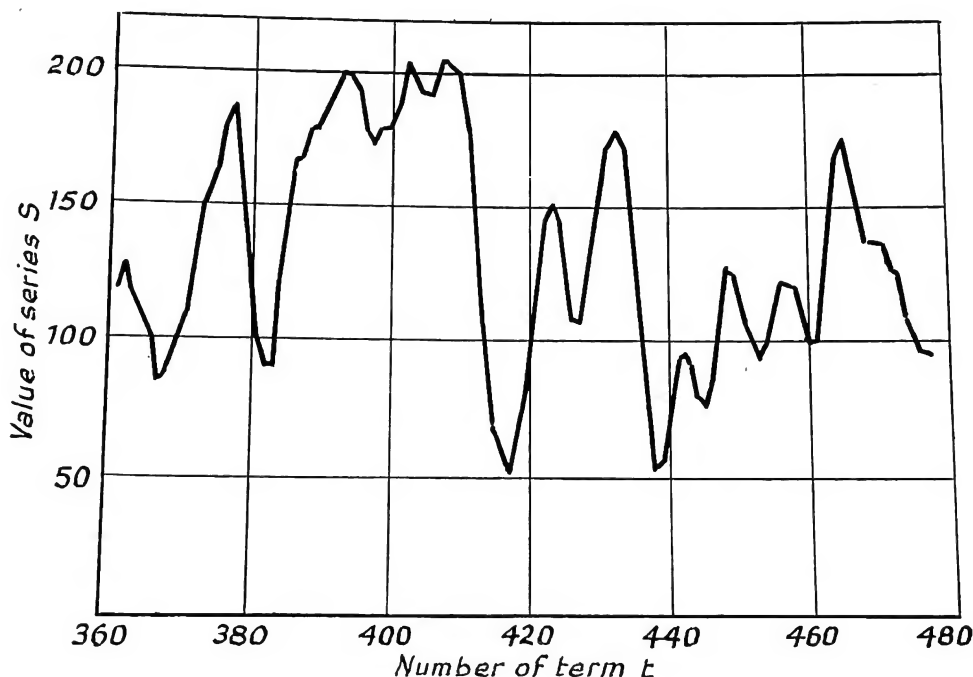


Fig. 46.1—Graph of the last 117 terms of the series S of Table 46.2

is 20.0 units, but this is based on rectangular variation, and we are, perhaps, entitled to expect some difference from normal theory. For rectangular random variables, values distant from the mean occur more frequently, and it is not surprising to find oscillations in the series which do not result in upcrosses.

The number of peaks in the series will be found to be 62, the first at the seventh term, the last at the 466th. Hence the mean distance between peaks is $459/61 = 7.5$ units. From formula (46.59) we find

$$\cos \theta_1 = \frac{2}{3}, \quad \theta_1 = 48^\circ 11'.$$

Thus the theoretical mean distance is $360/48.187 = 7.5$ units, in good agreement with experiment. It will be observed that several of the distances between peaks are due to very small ripples.

46.22 Let us now examine how the variance of the induced oscillation compares with the variance of the original random series.

The sum of k random elements with variance v has variance kv and its mean has variance v/k . It does not follow that a simple moving average has a variance $1/k$ times that of the random element, because of correlations between successive members in the derived series. If the original series was $\varepsilon_1, \dots, \varepsilon_n$, the derived series is, with weights a_1, \dots, a_k ,

$$\left. \begin{aligned} \sum a_j \varepsilon_j &= \eta_1, & \text{say} \\ \sum a_j \varepsilon_{j+1} &= \eta_2, & \text{say} \\ &\vdots & \\ \sum a_j \varepsilon_{j+n-k} &= \eta_{n-k+1}. \end{aligned} \right\} \quad (46.62)$$

The expected value of the sum of these values is zero since the expected value of ε may be taken to be so. Since there are $n-k+1$ terms we have for the variance

$$\frac{1}{n-k+1} \sum \eta^2. \quad (46.63)$$

The expected value of this, since the ε 's are independent, is

$$\frac{1}{n-k+1} E \sum \eta^2 = E(\eta^2) = v \sum_{i=1}^k a_i^2. \quad (46.64)$$

In particular, if the a 's are all equal to $1/k$, the expected value of the variance is v/k . This gives us the *average* reduction in the variance.

If a simple average of extent k is iterated q times the weights are the successive coefficients in

$$\frac{1}{k^q} (1 + x + x^2 + \dots + x^{k-1})^q = \frac{1}{k^q} \left(\frac{1-x^k}{1-x} \right)^q.$$

The sum of squares of these coefficients is the coefficient of $x^{q(k-1)}$ in

$$\frac{1}{k^{2q}} \frac{(1-x^k)^{2q}}{(1-x)^{2q}}, \quad (46.65)$$

and this gives the average reduced variance for a simple average of k iterated q times. The following are the values of the reducing factor for some of the values of k and q :

		q				
		1	2	3	4	5
k	3	0.33	0.23	0.19	0.17	0.15
	4	0.25	0.17	0.14	0.12	0.11
	5	0.20	0.14	0.11	0.10	0.09
	6	0.17	0.11	0.09	0.08	0.07
	7	0.14	0.10	0.08	0.07	0.06

Evidently the result of the first moving average is to generate a series with a much lower variance than that of the original random element, but the second and succeeding iterations do not reduce the variance further to the same extent. In the case $k = 7$ the first averaging reduces the variance to one-seventh, but the next three reduce it only by a further half.

46.23 It is also instructive to consider the effect of the moving average on serial correlations in residuals. For the series (46.51) generated by a moving average on a random series we have, as at (46.54),

$$\begin{aligned}\text{cov}(u_t, u_{t+s}) &= E\{\sum a_j \varepsilon_{j+t} \sum a_j \varepsilon_{j+s+t}\} \\ &= v \sum_{j=1}^{k-s} a_j a_{j+s},\end{aligned}\quad (46.66)$$

and thus for the s th serial correlation of the resultant series

$$r_s = \left. \begin{aligned} &\frac{\sum_{j=1}^{k-s} a_j a_{j+s}}{\sum_{j=1}^k a_j^2}, \quad |s| < k \\ &= 0, \quad |s| \geq k. \end{aligned} \right\} \quad (46.67)$$

Thus, for an infinite series generated in this way we see that, whereas the original (random) series had zero serial correlations, the induced series is serially correlated up to order k , i.e. as long as terms in the generated series have any terms of the original series in common.

For example with a simple moving average of extent k , all the a 's are equal to $1/k$, and from (46.67) we easily find

$$r_s = 1 - \frac{|s|}{k}, \quad (46.68)$$

so that the correlation may be quite high for $s = 1$ and falls off linearly, as s increases, to zero at $s = k$. High correlations of this kind between neighbouring values are responsible for the Slutsky-Yule effect.

Example 46.7

The weights of the Spencer 21-point formula are

$$\frac{1}{350}[-1, -3, -5, -5, -2, 6, 18, 33, 47, 57, 60].$$

Apart from the divisor 350, which may be disregarded for present purposes, the sum of squares of weights is 17,542. The products (46.66) and the corresponding serial correlations are as follows:

k	$\sum a_j a_{j+k}$	r_k	k	$\sum a_j a_{j+k}$	r_k
0	17,542	1.000	11	-930	-0.053
1	16,786	0.957	12	-528	-0.030
2	14,667	0.836	13	-214	-0.012
3	11,584	0.660	14	-27	-0.002
4	8,085	0.461	15	50	0.003
5	4,726	0.269	16	59	0.003
6	1,951	0.111	17	40	0.002
7	6	0.000	18	19	0.001
8	-1,074	-0.061	19	6	0.000
9	-1,430	-0.082	20	1	0.000
10	-1,298	-0.074	21	0	0.000

The correlogram is shown in Fig. 46.2. From $k = 13$ onwards the correlations are very small, and from $k = 21$ onwards they vanish completely.

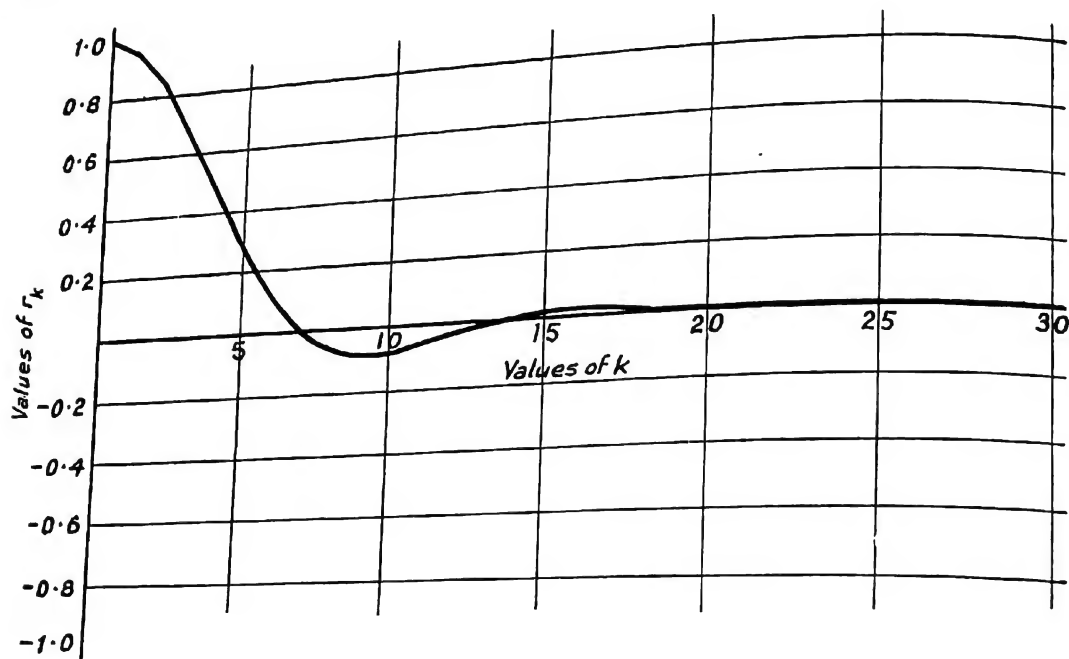


Fig. 46.2—Correlogram of series generated by the Spencer 21-point formula (Example 46.7)

The variate-difference method

46.24 The concept of a series which consists of a polynomial element plus a residual of a more or less random kind has given rise to a method which purports to eliminate the former by differencing. Clearly, successive differencing will eventually entirely eliminate any element which is actually a polynomial in the time, and may be relied upon almost to eliminate any systematic element except, perhaps, exponential or cyclical terms. Let us consider the effect of differencing upon a random series ε_t . We have

$$\begin{aligned}\Delta^r \varepsilon_t &= \varepsilon_{t+r} - \binom{r}{1} \varepsilon_{t+r-1} + \binom{r}{2} \varepsilon_{t+r-2} - \dots + (-1)^r \varepsilon_t \\ &= (U-1)^r \varepsilon_t.\end{aligned}\tag{46.69}$$

Taking, without loss of generality, ε_t to have zero mean, we have

$$E(\Delta^r \varepsilon_t) = 0,\tag{46.70}$$

and if ε_t has the same variance v for all t ,

$$\begin{aligned}\text{var}(\Delta^r \varepsilon_t) &= v \sum_{j=0}^r \binom{r}{j}^2 \\ &= v \times \text{coeff. } x^r \text{ in } (1+x)^r (x+1)^r \\ &= \binom{2r}{r} v.\end{aligned}\tag{46.71}$$

We may then derive an estimate of v by writing

$$v = \frac{\mu'_2(\Delta^r \varepsilon_t)}{\binom{2r}{r}}. \quad (46.72)$$

It is to be noticed that we use the second moment about zero, not the observed variance of $\Delta^r \varepsilon_t$, since the mean is known to be zero. This shortens the arithmetic to some extent.

The factor $\binom{2r}{r}$ for $r = 1$ to 10 has the following values:

r	$\binom{2r}{r}$	$1/\binom{2r}{r}$
1	2	0.5
2	6	0.166,667
3	20	0.05
4	70	0.014,285,7
5	252	0.003,968,25
6	924	0.001,082,25
7	3,432	0.000,291,375
8	12,870	0.000,077,700,1
9	48,620	0.000,020,567,7
10	184,756	0.000,005,412,54

46.25 Basing itself on equation (46.72), the method of variate differences proceeds as follows. We difference the series once, find the second moment about zero of the resultant, and divide by 2; we then difference again and find the second moment about zero, dividing in this case by 6; and so on. If the successive estimates of v decrease we continue with the differencing. There will, in general, come a point when they cease decreasing and remain constant within sampling limits (which may be rather wide). At this stage we may suppose that we have eliminated the systematic element in the original series. The final estimate gives us an estimate of the variance of the random element in the original series, and the order of the difference to which we have had to go will give an indication of the degree of the polynomial representing the systematic component.

Example 46.8

Let us apply the variate-difference technique to the series of Table 46.1. We know from the method of constructing the series that the systematic part ought to be completely eliminated after the third differencing, and also that the random part consists of an element with variance 833 approximately. In fact, the random numbers from 1 to N have a variance $(N^2 - 1)/12$, and N in this case is 100. The actual variance of the random element in Table 46.1 is 843.

THE ADVANCED THEORY OF STATISTICS

Table 46.3—Differences of the series u_t of Table 46.1

t	u_t	Δ^1	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6
1	-96	-6	67	155	279	508	1050
2	-90	-73	-88	-124	-229	-542	-1297
3	-17	15	36	105	313	755	1524
4	-32	-21	-69	-208	-442	-769	-1141
5	-11	48	139	234	327	372	271
6	-59	-91	-95	-93	-45	101	361
7	32	4	-2	-48	-146	-260	-229
8	28	6	46	98	114	-31	-625
9	22	-40	-52	-16	145	594	1661
10	62	12	-36	-161	-449	-1067	-2252
11	50	48	125	288	618	1185	1978
12	2	-77	-163	-330	-567	-793	-876
13	79	86	167	237	226	83	-159
14	-7	-81	-70	11	143	242	137
15	74	-11	-81	-132	-99	105	551
16	85	70	51	-33	-204	-446	-655
17	15	19	84	171	242	209	-64
18	-4	-65	-87	-71	33	273	690
19	61	22	-16	-104	-240	-417	-629
20	39	38	88	136	177	212	216
21	1	-50	-48	-41	-35	-4	175
22	51	-2	-7	-6	-31	-179	-650
23	53	5	-1	25	148	471	1110
24	48	6	-26	-123	-323	-639	-975
25	42	32	97	200	316	336	41
26	10	-65	-103	-116	-20	295	925
27	75	38	13	-96	-315	-630	-965
28	37	25	109	219	315	335	207
29	12	-84	-110	-96	-20	128	316
30	96	26	-14	-76	-148	-188	-32
31	70	40	62	72	40	-156	-798
32	30	-22	-10	32	196	642	1597
33	52	-12	-42	-164	-446	-955	-1719
34	64	30	122	282	509	764	950
35	34	-92	-160	-227	-225	-186	141
36	126	68	67	28	-69	-327	-991
37	58	1	39	97	258	664	1515
38	57	-38	-58	-161	-406	-851	-1492
39	95	20	103	245	445	641	707
40	75	-83	-142	-200	-196	-66	281
41	158	59	58	-4	-130	-347	-685
42	99	1	62	126	217	338	509
43	98	-61	-64	-91	-121	-171	-314
44	159	3	27	30	50	143	432
45	156	-24	-3	-20	-93	-289	-745
46	180	-21	17	73	196	456	
47	201	-38	-56	-123	-260		
48	239	18	67	137			
49	221	-49	-70				
50	270	21					
51	249						

Table 46.3 shows the series and the differences up to Δ^6 . For the sums of squares in the various columns S_j corresponding to Δ^j , we find:

$$\begin{aligned} S_1 &= 107,541 \\ S_2 &= 318,115 \\ S_3 &= 1,033,513 \\ S_4 &= 3,445,308 \\ S_5 &= 11,720,069 \\ S_6 &= 40,548,844 \end{aligned}$$

To obtain second moments we divide by $51-j$ and then, to obtain the estimate of v , by $\binom{2j}{j}$. We find the following:

j	Estimate
1	1075.41
2	1082.02
3	1076.58
4	1047.21
5	1011.05
6	975.20

Curiously enough, the estimate for $j = 2$ is higher than that for $j = 1$, and there is little difference between the various estimates. In the ordinary way we should have concluded that the systematic component was adequately represented by a polynomial of order 1, that is to say a straight line, and that the residual random element had a variance of about 1000.

The reader must not be surprised to find discrepancies of this kind between theory and experiment in short series; and the discrepancy is not, in fact, as big as it seems. The variance of the original series is 6272.61. The mean square of the first difference, divided by 2, is 1075.41, so that about five-sixths of the variance has been eliminated by the first differencing; and the method indicates, quite correctly, that the greater part of the systematic element is linear. The random element is rather large compared with the non-linear systematic terms, and the latter have got caught up in it—the series is too short for the variate-difference method to disentangle them. Consider, for instance, the cubic term $(t-26)^3/100$. In the original series this varies in value from -156.25 to $+156.25$. First differences reduce it to $3(t-26)^2/100$, varying from 18.75 through zero to 18.75, whereas the random element is increased in range from 0 to 198. Already the systematic term is being swamped by the random element, and a slight degree of accidental correlation between the two can easily account for the increase in the mean square of second differences.

The matter may be put in a slightly different way. Suppose that, relying on the variate-difference method, we regarded the data as represented by a linear equation plus a random residual. If we fitted a straight line by least squares and examined the residuals, we should probably find very little evidence of departure from randomness. This representation would differ from the mode of construction of the series,

but it would be a possible method of construction. Only the failure of the representation to conform to further terms of the series would reveal its weakness.

46.26 The variate-difference method thus provides a kind of lower limit to the degree of the polynomial which will represent a series locally or generally. There remains for consideration the question as to what sort of differences between successive estimates of v can be regarded as chance effects, in order to decide when the value has reached a stationary level. The sum of squares S_j is a constant factor times the second moment, but as its members are correlated among themselves we cannot use the variance of the second moment to test its significance. Further, S_j and S_{j+1} are correlated. We proceed to derive the sampling variance of their difference, the somewhat complicated formulae being due to O. Anderson (1914).

46.27 Write

$$b_j = \binom{r}{j}. \quad (46.73)$$

Then we have

$$\frac{E(\Delta^r u)^2}{\binom{2r}{r}} = \frac{(\sum b_j^2) \mu_2}{\sum b_j^2} = \mu_2, \quad (46.74)$$

where μ_2 is the variance of u . Further,

$$\begin{aligned} E(\Delta^r u)^4 &= E[\{b_0 u_{r+1} - b_1 u_r + b_2 u_{r-2} - \dots + (-1)^r b_r u_1\}^2 \\ &\quad + \{b_0 u_{r+2} - b_1 u_{r+1} + b_2 u_r - \dots + (-1)^r b_r u_2\}^2 \\ &\quad + \dots \\ &\quad + \{b_0 u_n - b_1 u_{n-1} + b_2 u_{n-2} - \dots + (-1)^r b_r u_{n-r}\}^2]^2. \end{aligned} \quad (46.75)$$

Consider first of all the terms in this which result in fourth powers of u . They will derive from

$$\begin{aligned} &E\{b_0^2 u_{r+1}^2 + b_1^2 u_r^2 + \dots + b_r^2 u_1^2 \\ &\quad + b_0^2 u_{r+2}^2 + b_1^2 u_{r+1}^2 + \dots + b_r^2 u_2^2 \\ &\quad + \dots \\ &\quad + b_0^2 u_n^2 + b_1^2 u_{n-1}^2 + \dots + b_r^2 u_{n-r}^2\}^2. \end{aligned} \quad (46.76)$$

Writing now

$$B_0^2 = (b_0^2)^2 + (b_0^2 + b_1^2)^2 + \dots + (b_0^2 + b_1^2 + \dots + b_{r-1}^2)^2 \quad (46.77)$$

$$A_0^2 = (b_0^2 + b_1^2 + \dots + b_r^2)^2 = \binom{2r}{r}^2, \quad (46.78)$$

we find that the term in $E(u^4)$ is

$$\{A_0^2(n-r) + 2B_0^2\} E(u^4). \quad (46.79)$$

The only other term appearing from (46.75) will be of type $E(u_l^2 u_m^2)$, $l \neq m$. If the reader will write out the expansion of (46.75) he will find that the coefficients are expressible in terms of

$$A_j^2 = (b_0 b_j + b_1 b_{j+1} + \dots + b_{r-j} b_r)^2 = \binom{2r}{r-j}^2 \quad (46.80)$$

and

$$B_j^2 = (b_0 b_j)^2 + (b_0 b_j + b_1 b_{j+1})^2 + (b_0 b_j + b_1 b_{j+1} + \dots + b_{r-j-1} b_{r-1})^2. \quad (46.81)$$

The expression for $E(\Delta^r u)^4$ reduces to

$$(n-2r)A_0^2 E(u^4) + 4\{(n-2r+1)A_1^2 + (n-2r+2)A_2^2 + \dots + A_r^2(n-2r+r)\} E(u_i^2 u_m^2) + 2B_0^2 E(u^4) + 8\{B_1^2 + B_2^2 + \dots + B_r^2\} E(u_i^2 u_m^2). \quad (46.82)$$

Substituting μ_4 for $E(u^4)$ and μ_2^2 for $E(u_i^2 u_m^2)$, dividing by $(n-r)^2 \binom{2r}{r}^2$ and subtracting μ_2^2 , we find the sampling variance of the estimate of v . The expression can, however, be simplified to some extent. Putting

$$T_r = \sum_{j=0}^{r-1} \binom{r}{j}^2 \binom{r}{j+1}^2 + 2 \sum_{j=0}^{r-2} \binom{r}{j}^2 \binom{r}{j+2}^2 + 3 \sum_{j=0}^{r-3} \binom{r}{j}^2 \binom{r}{j+3}^2 + \dots + r \binom{r}{0}^2 \binom{r}{r}^2, \quad (46.83)$$

we find, after lengthy algebraic rearrangement,

$$\text{var} \frac{S_r}{(n-r) \binom{2r}{r}} = \frac{\mu_4 - 3\mu_2^2}{n-r} \left\{ 1 - \frac{2T_r}{(n-r) \binom{2r}{r}^2} \right\} + \frac{2\mu_2^2}{n-r} \left\{ \frac{\binom{4r}{2r}}{\binom{2r}{r}^2} - \frac{r}{2(n-r)} \right\}, \quad r \leq \frac{1}{2}n. \quad (46.84)$$

If terms of order $(n-r)^{-2}$ can be neglected, this reduces to

$$\frac{\mu_4 - 3\mu_2^2}{n-r} + \frac{\binom{4r}{2r}}{\binom{2r}{r}^2} \cdot \frac{2\mu_2^2}{n-r} \quad (46.85)$$

or, using the Stirling approximation to factorials,

$$\frac{1}{n-r} \{ \mu_4 - 3\mu_2^2 + \mu_2^2 \sqrt{(2r\pi)} \}, \quad (46.86)$$

which is a fair approximation to (46.85), being within 3 per cent for r as low as 6.

When the population of values of u is normal, $\mu_4 - 3\mu_2^2$ vanishes and the formula simplifies accordingly.

46.28 In a similar way it may be shown that

$$\text{cov} \left\{ \frac{S_r}{(n-r) \binom{2r}{r}}, \frac{S_{r+1}}{(n-r-1) \binom{2r+2}{r+1}} \right\} = \frac{\mu_4 - 3\mu_2^2}{n-r} \left\{ 1 - \frac{2T'_r}{\binom{2r}{r} \binom{2r+2}{r+1} (n-r-1)} \right\} + \frac{2\mu_2^2}{n-r} \left\{ \frac{\binom{4r+1}{2r}}{\binom{2r}{r} \binom{2r+2}{r+1}} \frac{2n-2r-1}{n-r-1} - \frac{r+1}{2(n-r-1)} \right\}, \quad (46.87)$$

where

$$T'_r = \sum_{j=0}^{r-1} \binom{r}{j}^2 \binom{r+1}{j+2}^2 + 2 \sum_{j=0}^{r-2} \binom{r}{j}^2 \binom{r+1}{j+3}^2 + \dots + r \binom{r}{0}^2 \binom{r+1}{r+1}^2.$$

We can now determine the variance of the difference of

$$\frac{S_r}{(n-r) \binom{2r}{r}} \quad \text{and} \quad \frac{S_{r+1}}{(n-r-1) \binom{2r+2}{r+1}}.$$

The general formula is complicated, but for normal variation, large n and $r \geq 6$ we have, analogously to (46.86),

$$\text{var (difference)} = \frac{(3r+1) \sqrt{(2\pi r)}}{2(2r+1)^3 (n-r-1)} \left\{ \frac{S_r}{(n-r) \binom{2r}{r}} \right\}^2. \quad (46.88)$$

The arithmetic application of the formulae has been facilitated by the preparation of tables of the constants involved. Reference may be made to Tintner (1940) who gives tables prepared by himself, O. Anderson and Zaycoff. We shall consider below some further modifications which simplify the formulae to a certain extent.

Example 46.9

For the data of Table 45.4 (sheep population) an application of the variate-difference method up to the tenth difference gave the following results:

r	$S_r / \left(\binom{2r}{r} (n-r) \right)$
1	3468
2	1442
3	854
4	629
5	518
6	448
7	401
8	371
9	357
10	347

The values here are falling steadily from $r = 1$ to $r = 10$, but very slightly towards the end. From (46.88) for $r = 6$ we have for the variance of the difference, 80.7 approximately, and for $r = 10$, 25.8 approximately. It appears that the reduction in variance at $r = 8$ is losing significance. It does not, of course, follow that the trend-line must be of this degree, for we may not want to eliminate the oscillatory movements in the trend-line. In fact, we should not leave much behind if we eliminated trend by a cubic.

46.29 The variate-difference method will clearly not eliminate systematic effects such as periodic terms with very short period. Consider, for instance, the series 1, -1, 1, -1, etc. The first differences give us a series 2, -2, 2, -2, etc., second differences 4, -4, 4, -4, etc., and so on. The variance of the series of r th differences is, neglecting effects due to the shortness of the series, 2^{2r} times that of the original,

and the quotient when this is divided by $\binom{2r}{r}$ tends to

$$\frac{2^{2r}(r!)^2}{(2r)!} \sim \sqrt{(\pi r)}$$

and so increases without limit. In such a case we cannot obtain an estimate of the variance of any random element which may be present.

The problem of testing differences between S_r and S_{r+1} , or the equivalent problem of testing whether the ratio S_r/S_{r+1} is near unity, is complicated by correlations between the differences which compose these quantities. Tintner (1940) and Johnson (1948) have suggested methods of overcoming the difficulty, but they involve sacrificing a large proportion of the data.

46.30 There is an intimate connexion between the variance of the differences of a series and its serial correlations. We have for a series of n terms

$$\begin{aligned} \sum_{t=1}^{n-1} (\Delta u_t)^2 &= \sum_{t=1}^{n-1} (u_{t+1} - u_t)^2 = \sum_{t=1}^{n-1} \{(u_{t+1} - \bar{u}) - (u_t - \bar{u})\}^2 \\ &= \sum_{t=1}^{n-1} (u_{t+1} - \bar{u})^2 - 2 \sum_{t=1}^{n-1} (u_{t+1} - \bar{u})(u_t - \bar{u}) + \sum_{t=1}^{n-1} (u_t - \bar{u})^2. \end{aligned}$$

Approximately, then, on division by $n-1$,

$$\begin{aligned} \text{var } \Delta u_t &= \text{var } u - 2 \text{cov}(u_{t+1}, u_t) + \text{var } u \\ &= 2 \text{var } u(1 - r_1). \end{aligned} \quad (46.89)$$

To the same degree of approximation,

$$\begin{aligned} \text{var } \Delta^2 u_t &= \text{var } (u_{t+2} - 2u_{t+1} + u_t)^2 \\ &= \text{var } u(6 - 8r_1 + 2r_2). \end{aligned} \quad (46.90)$$

Likewise, we have in general

$$\begin{aligned} \text{var } \Delta^p u_t &= \text{var} \left\{ \sum_{j=0}^p \binom{p}{j} (-1)^j u_{t+j} \right\} \\ &= \text{var } u \left\{ \sum_{j=0}^p \binom{p}{j}^2 - 2r_1 \sum_{j=0}^{p-1} \binom{p}{j} \binom{p}{j+1} \right. \\ &\quad \left. + 2r_2 \sum_{j=0}^{p-2} \binom{p}{j} \binom{p}{j+2} - \dots \right\} \\ &= \text{var } u \left\{ \binom{2p}{p} - 2r_1 \binom{2p}{p-1} + 2r_2 \binom{2p}{p-2} - \dots \right\} \\ &= \binom{2p}{p} \text{var } u \left\{ 1 - \frac{2p}{p+1} r_1 + \frac{2p(p-1)}{(p+1)(p+2)} r_2 - \dots \right\}. \end{aligned} \quad (46.91)$$

We can similarly express the serial correlations in terms of variances of the differences. Put

$$V_j = \frac{\text{var } \Delta^j u}{\binom{2j}{j}}, \quad V_0 = \text{var } u. \quad (46.92)$$

Then it may be shown (cf. Exercise 46.14) that

$$V_0 r_p = V_0 - p^2 V_1 + \frac{p^2(p^2-1)}{(2!)^2} V_2 - \frac{p^2(p^2-1)(p^2-4)}{(3!)^2} V_3 + \dots \quad (46.93)$$

46.31 The relative simplicity of these formulae is due to the fact that we have dealt with a long series, i.e. have neglected certain end-effects, writing, for example,

$$\sum_{i=1}^{n-1} u_i^2 \doteq \sum_{i=1}^{n-1} u_{i+1}^2.$$

We can preserve the formulae as exact results if we are prepared to make some small modifications to the definitions so as to incorporate these end-effects from the outset.

Define a series of summation functions $\Sigma_{(i)}$ by the formulae

$$\Sigma_{(0)}(x_i y_i) = x_1 y_1 + x_2 y_2 + \dots + x_{n-1} y_{n-1} + x_n y_n \quad (46.94)$$

$$\Sigma_{(1)}(x_i y_i) = \frac{1}{2} x_1 y_1 + x_2 y_2 + \dots + x_{n-1} y_{n-1} + \frac{1}{2} x_n y_n \quad (46.95)$$

$$\Sigma_{(2)}(x_i y_i) = \frac{1}{4} x_1 y_1 + \frac{3}{4} x_2 y_2 + x_3 y_3 + \dots + x_{n-2} y_{n-2} + \frac{3}{4} x_{n-1} y_{n-1} + \frac{1}{4} x_n y_n. \quad (46.96)$$

The general law of formation obeys the recurrence rule

$$\sum_{i=1}^n \Sigma_{(m)}(x_i y_i) = \frac{1}{2} \sum_{i=1}^{n-1} \Sigma_{(m-1)}(x_i y_i + x_{i+1} y_{i+1}), \quad (46.97)$$

so that, for example, the first three terms in $\Sigma_{(3)}$ have coefficients $\frac{1}{8}, \frac{1}{2}, \frac{7}{8}$.

Now define the modified quantities

$${}_m V_0 = \Sigma_{(m)} u^2 / n \quad (46.98)$$

$${}_m V_1 = \Sigma_{(m-1)} (\Delta u)^2 / 2n \quad (46.99)$$

$${}_m V_p = \Sigma_{(m-p)} (\Delta^p u)^2 / \binom{2p}{p} n. \quad (46.100)$$

Likewise define

$${}_m r_p = \Sigma_{(m-p)} u_i u_{i+p} / \Sigma_{(m)} u_i^2. \quad (46.101)$$

Then these quantities obey exactly the relations (46.91) and (46.93). The simplest way of seeing this, perhaps, is to consider the series

$$0, 0, 0, u_1, u_2, \dots, u_n, 0, 0, 0, \dots \quad (46.102)$$

The first differences are

$$0, 0, u_1, u_2 - u_1, \dots, u_n - u_{n-1}, -u_n, 0, 0, \dots \quad (46.103)$$

and their sum of squares is

$$u_1^2 + \sum_{j=1}^{n-1} (u_{j+1} - u_j)^2 + u_n^2 = 2 \sum_{j=1}^n u_j^2 - 2 \sum_{j=1}^{n-1} u_j u_{j+1} = 2 {}_1 V_0 (1 - {}_1 r_1). \quad (46.104)$$

In fact, for such a series, (46.97) is equivalent to

$$\begin{aligned} \sum_{i=1}^n \Sigma_{(m)} u_i u_{i+k} &= \frac{1}{2} \sum_{i=1}^n \Sigma_{(m-1)} (u_i u_{i+k} + u_{i+1} u_{i+1+k}) \\ &= (\Sigma_{(m)} u_i^2) ({}_m r_k) \end{aligned}$$

and the argument by which we arrived at (46.91) and (46.93) holds exactly for the infinite series, hence for the infinite series (46.102) and hence for our modified V 's and r 's.

46.32 As we have approached the matter, the variate-difference method has been used primarily to examine the order of the polynomial of best fit, a point being reached when the quantities V do not seriously change for higher differences. But on the assumption that the original series consisted of a polynomial plus a random

error we can also enquire, given a set of V 's, what is the best estimate of the error variance. The question has been examined by Quenouille (1953b), who seeks a linear function of the V 's which has minimal variance. In most practical cases it is more realistic to consider the possibility of serial correlation among the errors. Given a series consisting of a polynomial plus a serially correlated error, the problem is then to extend the variate-difference method so as to estimate the serial correlations. This question was also considered by Quenouille. Cf. Exercises 46.9–11.

46.33 But, however we approach the subject—by fitting polynomials directly, by moving averages, or by any other smoothing process—we encounter the difficulties mentioned in 46.18. The trend elimination will distort the residuals. There seems no escape from this situation. We can only hope to make the best of it, and this we can do in two ways: by choosing methods which, other things being equal, minimize the distortion; and by arranging our procedure so that, if we have misgivings at any stage of the later analysis, we can disentangle the distortion due to smoothing from other elements in the residual series. We proceed to examine the possibilities of the second line of attack.

46.34 Let us suppose that we divide our series of n terms into consecutive sets of s , and fit a polynomial of the same type to each set. Within any one set we may get a satisfactory trend-line. But clearly the line for any set must join on to the line for the next, and, in some acceptable sense, smoothly so. Subject to this matter, which we shall examine in a moment, such a method has the advantage that it treats the series as a set of independent blocks, and we can apply an analysis of variance to them. The method may be regarded as a compromise between the moving average and fitting a polynomial to the whole series. It was considered at length by Rhodes (1921) and has been extended by Quenouille (1949a).

As a simple example of the method, consider the fitting of straight lines to sets of three points. If the fitted value at u_1 is $2b_1$, and u_3 is $2b_2$, the value at u_2 must be $b_1 + b_2$. If, further, the fitted value to u_5 is $2b_3$ (that at u_3 being already determined as $2b_2$), the value at u_4 is $b_2 + b_3$, and so on, the values being

actual	u_1	u_2	u_3	u_4	u_5	u_6	u_7	\dots	
fitted	$2b_1$	$b_1 + b_2$	$2b_2$	$b_2 + b_3$	$2b_3$	$b_3 + b_4$	$2b_4$	\dots	(46.105)

The trend-line so determined will be continuous, but its first derivative will be discontinuous at u_3, u_5, u_7 , etc., i.e. it consists of a series of straight lines of extent three.

46.35 The actual values of the constants b may be determined in the usual way by least squares, i.e. we may minimize

$$(u_1 - 2b_1)^2 + \{u_2 - (b_1 + b_2)\}^2 + (u_3 - 2b_2)^2 + \text{etc.}, \quad (46.106)$$

giving the set of equations

$$\begin{aligned} 2u_1 + u_2 - 5b_1 - b_2 &= 0 \\ u_2 + 2u_3 + u_4 - b_1 - 6b_2 - b_3 &= 0 \\ \text{etc.} \end{aligned} \quad (46.107)$$

The equations are not difficult to solve, but once again they can be simplified very

394 THE ADVANCED THEORY OF

much by a suitable modification of end-effects. Let us, as usual, consider an odd number of terms u_1, \dots, u_{2m+1} and modify our series to

$$\frac{1}{2}(u_1 + u_{2m+1}), u_2, u_3, \dots, u_{2m}, \frac{1}{2}(u_1 + u_{2m+1}) \quad (46.108)$$

$$2b, b, + b_m. \quad (46.109)$$

This is analogous to the procedure of 45.34, in which we take the last term as equal to the first and hence render the series "circular." Writing u'_1 for $\frac{1}{2}(u_1 + u_{2m+1})$, we have to minimize the sum

$$\{u'_1 - (b_1 + b_m)\}^2 + (u_2 - 2b_1)^2 + \dots + (u_{2m} - 2b_m)^2$$

$$\left. \begin{array}{l} + b_m = u'_1 + 2u_2 + u_3 = U_2, \text{ say} \\ = u_3 + 2u_4 + u_5 = U_4, \text{ say} \\ = u_5 + 2u_6 + u_7 = U_6, \text{ say} \\ \\ + b_{m-1} + 6b_m = u_{2m-1} + 2u_m + u'_1 = U_{2m}, \text{ say.} \end{array} \right\} \quad (46.110)$$

The advantage of this form is that the coefficients of the b 's form a symmetrical circulant matrix which can be solved once and for all. (For the method see Quenouille, 1949a, and Good, 1950—cf. Exercise 46.13.) The result is to express the b 's in terms of the linear functions of the U 's in (46.110).

Exercise 46.12 generalizes the results above.

Example 46.10

We revert to the data of Table 46.1 with values of u_i as given in column (4), except that $\frac{1}{2}(u_1 + u_{51})$ has been substituted for u_1 and u_{51} . The values are repeated in Table 46.4. The functions U_i are shown in column (3) and the corresponding values of b in column (4). Thus, for example,

$$U_2 = 76.5 + 2(-90) + (-17) = -120.5$$

$$U_{50} = 221 + 2(270) + 76.5 = 837.5.$$

Also

$$\begin{aligned} 6b_1 + b_2 + b_m &= -242.2086 - 5.6002 + 127.3086 \\ &= -120.5002 = U_2, \end{aligned}$$

and so on. The fitted values are immediately obtainable, e.g. for $t = 2$, $u_2 = -90$, $2b_1 = -80.74$; for $t = 3$, $u_3 = -17$, $b_1 + b_2 = -45.97$. For $t = 1$, $u_1 = 76.5$, $b_1 + b_{25} = 86.94$.

As usual in Least Squares fitting, we do not need to work out each residual in order to calculate the sum of squares, for (cf. (35.1)) the Residual SS is

$$\sum u_t^2 - \sum b_i U_i.$$

In our present example

$$\sum u_i^2 = 474,458.25$$

$$\sum b_i U_i = 448,274.26$$

Difference = 26,184.

Difference = 26,184.

We have fitted 25 constants to 51 observations, one of which was adjusted, so there are

Table 46.4—Series of Table 46.1 fitted by straight lines to sets of threes

t	u_t	U_t	b_t	t	u_t	U_t	b_t
1	76.5			27	75		
2	-90	-120.5	-40.3681	28	37	161	17.6740
3	-17			29	12		
4	-32	-92	-5.6002	30	96	274	39.4348
5	-11			31	70		
6	-59	-97	-18.0309	32	30	182	19.7171
7	32			33	52		
8	28	110	16.7855	34	64	214	24.2620
9	22			35	34		
10	62	196	27.3178	36	126	344	48.7106
11	50			37	58		
12	2	133	15.3071	38	57	267	27.4740
13	79			39	95		
14	-7	139	13.8390	40	75	403	53.4452
15	74			41	158		
16	85	259	40.6586	42	99	454	54.8549
17	15			43	98		
18	-4	68	1.2096	44	159	572	71.4253
19	61			45	156		
20	39	140	20.0840	46	180	717	88.5930
21	1			47	201		
22	51	156	18.2862	48	239	900	114.0164
23	53			49	221		
24	48	191	26.1987	50	270	837.5	127.3086
25	42			51	76.5		
26	10	137	15.5212				
				TOTALS	-	6545.0	818.1244

25 degrees of freedom in the estimate of residual variance. The estimator is then $26,184/25 = 1047$ against a value obtained (from first differences) in Example 46.9 of 1075.

But we can do better than this. The method of fitting lines to three points suggests that there may be correlation between observed residuals in neighbouring points of a set of three, but not between sets.

We can, in fact, regard the series as $\frac{1}{2}(n-1)$ blocks of two, the two differences in a fitted triad having values typified by $b_2 - b_1$, $b_2 - b_1$. The sum of squares within blocks is estimated by $\frac{1}{50}(\sum u_i - \sum u_{i+1})^2$ which is found to be 406.12. Thus we can analyse the total sum of squares as

	d.fr.	SS	Mean square
Fitted constants	25	448,274.26	
Within blocks	1	406.12	
Residual	24	25,777.87	1074
	50	474,458.25	

The residual mean square is now in almost exact agreement with the value obtained

in Example 46.9 from first differences. In fact, the agreement is much closer than we have any right to expect.

For a more extended account of this topic, reference should be made to Quenouille (1949a) who has subsequently considered the problem of fitting so that the pieces of trend-line join smoothly.

Seasonal variation

46.36 In an analysis of seasonal variation we are presented with one major advantage and one minor disadvantage. The advantage is that we know the period of the seasonal recurrence to be one year. The disadvantage is that our observations are usually at an even number of points, quarters, months, or weeks. Most of what we have to say about seasonal movements over a year can be applied without change to movements over other periods which are strictly cyclical, e.g. temperature movements over a day or price movements over a week. For simplicity, however, we shall confine ourselves to seasonality in the strict sense of the word.

46.37 Seasonal movements are often sufficiently marked to need no demonstration. Cases sometimes occur, however, where we are not certain whether the movements in a series are due to random effects imposed on a trend or to a fluctuation of non-cyclical character, and in the first instance we require a test for the existence of seasonality. In any case we require a measure of the seasonal effects.

The quarterly data of Table 46.5 (index-numbers of wholesale prices of vegetable food) will illustrate the possibilities. In Table 46.6 we have simplified the data by taking a new origin and scale.

Table 46.5—Quarterly index numbers of the wholesale price of vegetable food in the United Kingdom, 1951–8

(Data from the *Journal of the Royal Statistical Society* for appropriate years. 1867–1877 = 100)

	1951	1952	1953	1954	1955	1956	1957	1958
First quarter	295.0	324.7	372.9	354.0	333.7	323.2	304.3	312.5
2nd "	317.5	323.7	380.9	345.7	323.9	342.9	285.9	336.1
3rd "	314.9	322.5	353.0	319.5	312.8	300.3	292.3	295.5
4th "	321.4	332.9	348.9	317.6	310.2	309.8	298.7	318.4

Table 46.6—Data of Table 46.5 with origin 300, values multiplied by 10

	1951	1952	1953	1954	1955	1956	1957	1958	Totals	Mean
First quarter	-50	247	729	540	337	232	43	125	2203	275.375
2nd "	175	237	809	457	239	429	-141	361	2566	320.750
3rd "	149	225	530	195	128	3	-77	-45	1108	138.500
4th "	214	329	489	176	102	98	-13	184	1579	197.375
TOTALS	488	1038	2557	1368	806	762	-188	625	7456	

46.38 Consider first of all the possibilities of distribution-free tests. It is tempting to rank the quarters within any one year from 1 to 4 and consider how the ranks vary from year to year. A little reflection will show, however, that such a procedure does not disentangle seasonal movement from trend. If the data were uniformly increasing in time the first quarter would always rank the lowest; but this is not a seasonal effect. In fact, to make any progress it appears necessary to make some attempt to eliminate trend as a first step.

One simple approach is to assume that deviations from the annual average are seasonal. This is equivalent to supposing that there is a trend from year to year but that, within a year, departures from the year's average are seasonal effects. This may obviously be a somewhat indifferent approximation to the truth. If we are prepared to adopt it, the procedure in Table 46.6 would be as follows.

Over the eight years concerned the means for the four quarters are 275.375, 320.750, 138.500, 197.375, themselves with a mean of 233.000. The deviations from this mean are then 42.375, 87.750, -94.500, -35.625. In terms of the original variables of Table 46.5 the corresponding values would be 304.2375, 308.7750, 290.5500, 296.4375, or, on the basis of a mean of 100, a third of these values, namely 101.41, 102.93, 96.85, 98.81. These are taken as indexes of seasonal variation. The general procedure, to eliminate seasonality from the original data, would then be to divide all the first-quarter figures by 101.41, all the second-quarter figures by 102.93, and so on.

To test these seasonal indexes, we must write down the model. Our present procedure is to assume that each observation is the sum of three effects: a yearly value, say y , a seasonal value (constant from year to year in proportional effect), say s , and an error which is random. Thus,

$$u_t = y_t s_q + \varepsilon, \quad t = 1, \dots, n; \quad q = 1, 2, 3, 4. \quad (46.111)$$

If the trend is slow, so that the seasonal effect may be regarded as constant from year to year in absolute (not proportional) magnitude, we can write approximately

$$u = y_t + s_q + \varepsilon, \quad (46.112)$$

which is an ordinary analysis of variance model with a two-fold classification. If the trend is not slow we ought to work with logarithms so as to obtain

$$\log u = \log y + \log s_q + \eta. \quad (46.113)$$

The weakness of this model is apparent. We assume that y is a constant for the year so that, for example, the value for the fourth quarter of the first year is $y_1 + s_4$, and that for the succeeding quarter is $y_2 + s_1$. These values may not "join on" smoothly in the way required by our intuitive feeling about the smoothness of trend. We shall therefore not bother with the arithmetic analysis. Indeed, we should have dismissed this method more summarily were it not for the fact that it is widely used in elementary texts.

46.39 A second possibility is to use a moving average to eliminate trend before examining the residual values for seasonality. We then, of course, run into the danger of distorting the residuals. However, if we choose our moving average with care, we can minimize this effect so far as concerns seasonal effects. We noted, in fact, in 46.16 that if the simple moving average (with equal weights) is equal in extent to

Table 46.7—Residuals in data of Table 46.6 after elimination of trend by a centred moving average of fours

	1951	1952	1953	1954	1955	1956	1957	1958	Totals	Means
First quarter										
2nd		25.750	167.875	77.875	108.625	24.875	52.250	22.000	479.250	68.46
3rd	-10.125	-8.125	189.750	75.875	28.250	238.000	107.875	229.375	645.250	92.18
4th	10.000	-94.750	-85.625	-121.625	-60.375	-163.875	-40.250		-576.625	-82.38
		-122.500	-59.000	-88.000	-97.000	26.000	-49.250		-379.750	-54.25

Table 46.8—Residuals in data of Table 46.6 after elimination of trend by a cubic fitted to seven points

	1951	1952	1953	1954	1955	1956	1957	1958	Totals	Means
First quarter										
2nd		+30.38	122.14	79.95	113.19	16.33	91.14		458.56	65.51
3rd		+29.19	135.62	82.24	27.43	206.57	-84.24	5.43	396.81	66.13
4th	-12.67	-53.71	-123.95	-106.81	-35.76	-191.57	12.38		-499.42	-83.24
		-128.67	-72.57	-70.48	-97.90	25.81	-54.00		-410.48	-58.64

the period of a cyclical component, the trend-value of that component is zero, so that the residual is unimpaired.

For the data of Table 46.5, this involves the use of a simple moving average of fours, which will be adequate to remove a linear trend. Our original data, however, are the averages for a quarter's prices and relate, then, to time periods of three months centred on February 15, May 15, August 15, November 15 (or thereabouts). The average of an (even) set of four will give us a trend-value at some point half-way between these dates. To bring the time-point of the average back to comparability with the originals we must "centre" the average. This is most simply carried out by taking the mean of consecutive pairs of the four-point average. Thus, in Table 46.6 the mean of the first four values is 122, and of the next four values is 196.25. The mean of these two, 159.125, is taken as the trend-value corresponding to the third quarter of 1951, namely where the original value of the series is 149. The process is clearly equivalent to fitting a five-point average with weights

$$\frac{1}{8}[1, 2, 2, 2, 1]. \quad (46.114)$$

Proceeding in this manner on the data of Table 46.6, we find the residuals shown in Table 46.7. The deviations of the means (each based on seven months) from the overall mean of 24.05/4 are 62.45, 86.17, -88.39, -60.26. In terms of the original variables the corresponding values would be 306.245, 308.617, 291.161, 293.974 or, on the basis of a mean of 100, 102.08, 102.87, 97.06, 97.99. These are substantially different from the results of the method of 46.38.

46.40 It is also of interest to see what happens if we eliminate trend by a more elaborate form of moving average, and we will consider the fitting of a cubic to seven points with weights $\frac{1}{21}[-2, 3, 6, 7]$. The residuals are shown in Table 46.8. The seasonal indexes will be found to be

$$102.27, \quad 102.29, \quad 97.31, \quad 98.13$$

as compared with

$$101.41, \quad 102.93, \quad 96.85, \quad 98.81 \quad (\text{in } 46.38)$$

$$102.08, \quad 102.87, \quad 97.06, \quad 97.99 \quad (\text{in } 46.39)$$

Although the general picture is the same in all cases (a seasonal peak in the second quarter, a seasonal trough in the third) there are large enough differences in these results to embarrass us in work requiring great accuracy. Our inclination would be to use the method of 46.39. That of 46.40 runs into some danger of fitting too well, in the sense that the trend-line may embody some part of the seasonal effect. It seems impossible, however, to lay down any completely objective rules for the treatment of seasonal effect versus trend. Our general recommendation would be to try several methods and to choose the one which appears to give the most reasonable results; and, in any published work, to state exactly what has been done.

46.41 From the point of view of spectrum analysis, which we discuss in Chapter 49, there is more to be said about the effect of trend elimination both on random residuals and on seasonal components. We shall see that it is possible to correct the power spectrum for distortions due to trend fitting, at least in certain cases.

46.42 There are now in existence some complicated routines written for electronic computers which dissect a series into trend, seasonal and residual components. They may involve more than one process for the isolation of each component, for example by a preliminary smoothing, a first approximation to seasonals, a more refined trend fitting, a further approximation to seasonals, and so on. The proof of all these puddings is in the eating, and it seems fair to say that for a wide class of economic and social statistics such routines work quite well in practice. Reference may be made to Shiskin (1955), Eisenpress (1956), Shiskin and Eisenpress (1957), and Burman (1965) for some work on this subject.

EXERCISES

46.1 A straight line is fitted to $2m+1$ points at equidistant unit intervals $-m, \dots, m$. Show that the line is

$$\frac{1}{2m+1} \sum u_t + \frac{3}{m(m+1)(2m+1)} \sum tu_t.$$

Hence show that the sum of squares of coefficients of a moving average based on this line for the point $t = j$ is

$$\frac{1}{2m+1} \left[1 + \frac{3j^2}{m(m+1)} \right].$$

46.2 Fit a cubic to the last seven points of the sheep series of Table 45.4 and show that it gives a trend for the final four values of 1639, 1687, 1750 and 1807.

46.3 Show that the weights in the Spencer 21-point formula are

$$\frac{1}{350} [-1, -3, -5, -5, 6, 18, 13, 47, 57, 60]$$

and that if it is applied to a random series the variance of the resultant is about one-seventh of the original series—about the same reduction as would be given by a simple moving average of sevens.

46.4 Show that Macaulay's 43-point formula

$$\frac{1}{960} [12] [8] [5]^2 \left[\frac{7}{10}, -1, 0, 0, 0, 0, 0, 0, 1 \right]$$

has weights

$$\frac{1}{9600} [7, 18, 30, 40, 45, 28, -8, -60, -122, -178, -205, -190,$$

$$-127, -6, 163, 360, 562, 760, 928, 1050, 1127, 1156]$$

and that it reduces the variance of a random series about as much as a simple average of nines.

46.5 If ε_t is a random series, show that the correlation between successive members of $\Delta^k \varepsilon_t$ for long series is $-k/(k+1)$ and hence tends to -1 as k increases. Hence show that the signs of successive terms in $\Delta^k u_t$ tend to alternate, where u_t is the sum of a random element and a systematic element representable by a polynomial; and verify by reference to Table 46.1.

46.6 By eliminating δ^2 from (46.28), show that, for a cubic curve, an accurate trend-line

is given by

$$\frac{1}{h^2 - k^2} \left\{ \frac{h^2 - 1}{k} [k] - \frac{k^2 - 1}{h} [h] \right\}$$

and generalize this result.

(Cf. Higham, 1882-5)

46.7 Show that in a long random series of normal elements with variance σ^2 the serial correlations are uncorrelated, and that

$$\text{var } r_k = 1/n.$$

Hence from (46.92) derive the large-sample formula

$$\text{var } V_p = \frac{2\sigma^4}{n} \left\{ 1 + \frac{2p^2}{(p+1)^2} + \frac{2p^2(p-1)^2}{(p+1)^2(p+2)^2} \dots \right\}. \quad (\text{Quenouille, 1953b})$$

46.8 Show that in the notation of 46.30 as applied to a random series with variance σ^2 and fourth cumulant κ_4 ,

$$\text{cov}(V_i, V_j) - \frac{\kappa_4}{n} = \frac{\sigma^4}{2n} \frac{\binom{2i+2j}{i+j}}{\binom{2i}{i} \binom{2j}{j}} = \frac{\sigma^4}{n} a_{ij}, \quad \text{say}. \quad (\text{Quenouille, 1953b})$$

46.9 In the previous exercise, given an estimator of the error variance

$$t = \sum_{i=m+1}^{m+p} c_i V_i$$

with variance $(\kappa_4 + \lambda\sigma^4)/n$, show that t has minimal variance if

$$\sum_{i=m+1}^{m+p} a_{ij} c_i - \lambda = 0, \quad j = m+1, \dots, m+p,$$

$$-\sum_{i=m+1}^{m+p} c_i = -1.$$

(Quenouille, 1953b)

46.10 A series consists of a polynomial in t of degree m plus a component which has the same variance for all t but the successive values of which may be serially correlated.

Define

$$R_{1p} = \frac{(p+1)(p+2)}{2!} \Delta V_p, \quad p = m, m+1, \dots$$

Show that for a long series

$$R_{1p} = V_0 \left\{ r_1 - \frac{4p}{p+3} r_2 + \frac{9p(p-1)}{(p+3)(p+4)} r_3 - \frac{16p(p-1)(p-2)}{(p+3)(p+4)(p+5)} r_4 + \dots \right\}$$

and hence that if the serial correlations higher than the first order vanish,

$$E(R_{1p}) = \sigma^2 r_1.$$

46.11 Continuing the previous exercise, show that, defining

$$R_{mp} = \frac{(p+2m-1)(p+2m)}{(2m-1)(2m)} \Delta R_{m-1, p}$$

and if serial correlations higher than the m th vanish, that

$$E(R_{mp}) = \sigma^2 r_m.$$

(Quenouille, 1953b)

46.12 In generalization of 46.34 show that if constants a are defined by
 $(1+t+t^2+\dots+t^{s-1})^{d+1} = a_0 + a_1 t + a_2 t^2 + \dots$

then the sets of weights $\sum a_{is} b_{i+1}, \sum a_{is-1} b_{i+1}, \sum a_{is-2} b_{i+1}$

fit a curve of degree d to sets of s points.

(Quenouille, 1949a)

46.13 If a circulant matrix A is given by

$$A = \{a_{j-i}\}, \quad i, j = 0, 1, \dots, n-1,$$

where the suffixes are reduced to modulus n , show that

$$|A| = \prod_{s=0}^{n-1} \sum_{r=0}^{n-1} a_r \omega^{rs}$$

where ω is $\exp(2\pi i/n)$, an n th root of -1 . Show also that the latent vectors λ of $|A|$ are given by

$$\lambda_s = \sum_r a_r \omega^{rs}, \quad s = 0, 1, \dots, n-1.$$

Show also that

$$a_s = \frac{1}{n} \sum_{r=0}^{n-1} \lambda_r \bar{\omega}^{rs}.$$

Hence, noting that the latent roots of A^{-1} are λ^{-1} , show how to invert A .

(Good, 1950)

46.14 Starting from the relation

$$(-1)^p \cos 2p\theta = 1 - \frac{p^2}{2!} (2 \cos \theta)^2 + \frac{p^2(p^2-1)}{4!} (2 \cos \theta)^4 - \dots$$

and putting $z = e^{i\theta}$, derive equation (46.93).

46.15 A cubic $\sum_{j=0}^s a_j t^j$ is fitted by least squares to a set of values $t = -m, \dots, 0, \dots, m$, with $n = 2m+1$. Show that it is given by

$$\begin{aligned} & \frac{15}{n(n^2-1)(n^2-4)} \left\{ \frac{1}{20} (n^2-1)(3n^2-7) \sum u - (n^2-1) \sum t^2 u \right\} \\ & + \frac{140t}{n(n^2-1)(n^2-4)(n^2-9)} \left\{ \frac{5}{28} (3n^4-18n^2+31) \sum tu - (3n^2-7) \sum t^3 u \right\} \\ & + \frac{15t^2}{n(n^2-1)(n^2-4)} \{ -(n^2-1) \sum u + 12 \sum t^2 u \} \\ & + \frac{140t^3}{n(n^2-1)(n^2-4)(n^2-9)} \{ -(3n^2-7) \sum tu + 20 \sum t^3 u \}. \end{aligned}$$

46.16 A random series has trend "eliminated" by the removal of a moving average with weights $[a_{-m}, a_{-(m-1)}, \dots, a_0]$. Show that the serial correlation of residuals is given approximately by

$$r_k = \frac{\sum_{i=-m}^{m-k} a_i a_{i+k} - 2a_k}{\sum_{i=-m}^m a_i^2 - 2a_0 + 1}.$$

Compare the values given by applying the formula to the residuals of Table 46.1 (col. (10)) with the actual values $r_1 = -0.411$, $r_2 = -0.244$, $r_3 = 0.231$, $r_4 = 0.143$, $r_5 = 0.007$.

CHAPTER 47

STATIONARY TIME-SERIES

47.1 If we remove from a time-series the elements attributable to seasonal variation and trend we shall, in general, be left with a series oscillating about some constant value. This movement may be so small as to be virtually non-existent—the series then consists entirely of seasonality or trend. Or the seasonality or trend may themselves be non-existent, in which case the series is entirely oscillatory. In the present chapter we shall study these oscillatory series, supposing that trend and seasonal effects have been eliminated or do not exist. Strictly speaking, we ought, perhaps, to treat seasonal effects as part of the oscillatory movement and not regard them as eliminated beforehand. But we shall see that there are types of oscillation (the rule rather than the exception) which are not seasonal in our sense, and it is better to keep them distinct as far as possible.

47.2 Let us begin with some intuitive ideas. In Table 45.1 (barley yields) we have an example of a series which fluctuates about a mean value to about the same extent. But we might have a series, as in Fig. 47.1(a), in which the extent of oscillation systematically increased or, as in Fig. 47.1(b), in which the amount of oscillation itself oscillated. We shall exclude such cases from discussion and confine our attention

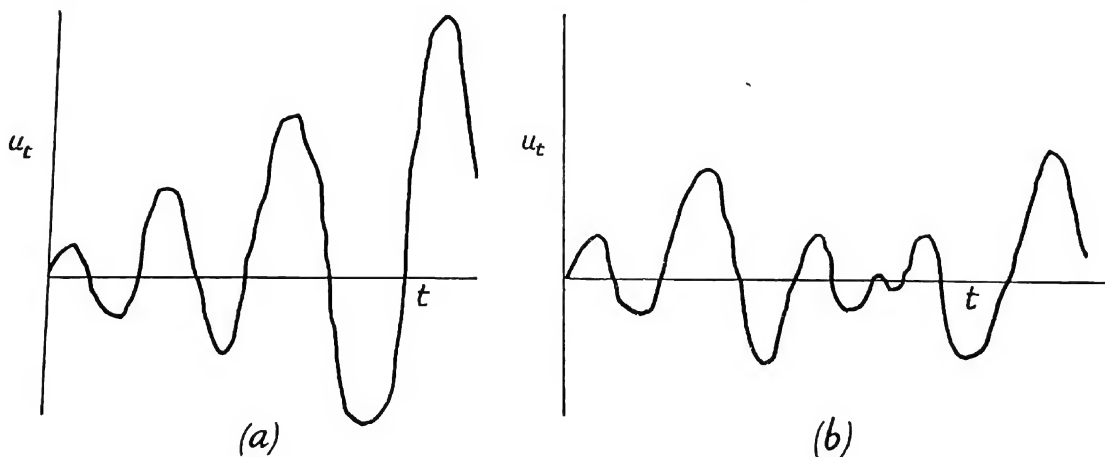


Fig. 47.1 (see text)

to series for which the amplitude remains more or less constant. This does not mean that the amplitude of the swings has to be exactly the same, but that there is no systematic effect present.

47.3 To make these ideas precise, consider a set of random variables arranged in order: $u_1, u_2, \dots, u_n, \dots$. Let the distribution function of any set of n consecutive u 's, say $u_{t+1}, u_{t+2}, \dots, u_{t+n}$, be

$$F(u_{t+1}, u_{t+2}, \dots, u_{t+n}). \quad (47.1)$$

Then if F is independent of t for all integral $n > 0$ we shall say that F represents a stationary series. The distribution of any set of n consecutive variables is the same, wherever in the series we choose it. In particular, for $n = 1$, we see that the distributions of all members of the series are identical.

In the theory of stochastic processes, of which stationary time-series are a particular case, it is customary to define stationarity of a less restrictive kind, e.g., a process is stationary *in the mean* if the expected values of all the u 's are the same, and it is stationary *in the variance* if all u 's have the same variance. Most of the applications of the general stationarity property (47.1) which we shall make concern the constancy of mean and variance along the series, but important use will also be made of product-moments of the u 's and of the identity of distributions of the u 's.

47.4 In regarding a sequence of variables of unlimited extent as defined by a distribution function we arrive at a new problem in the definition of mean values. We can, first of all, consider the behaviour of some u , say u_1 , for different series generated by (47.1). Or, in the second place, we can consider the limit of some set of u 's (or a function of them), say u_1, u_2, \dots, u_n as n tends to infinity.

In the first class of case we have to consider averages in a population composed of different ways in which the series could happen. Each such series is possible, and any one which occurs is called a *realization* of the process. We may have only one realization to examine, and, indeed, this is the rule rather than the exception. In a sense, then, we have only a sample of one observation from the process. We shall see shortly why this does not seriously limit the possibility of inferences about the process.

47.5 We shall, from now on, assume that the mean and variance of u exist. We then have, for all t ,

$$\mu = E(u_t) = \int_{-\infty}^{\infty} u_t dF(u_t) \quad (47.2)$$

$$\sigma^2 = E(u_t - \mu)^2 = \int_{-\infty}^{\infty} (u_t - \mu)^2 dF(u_t). \quad (47.3)$$

We also assume that any pair of u 's have an *autocovariance*

$$\gamma_j = E\{(u_t - \mu)(u_{t+j} - \mu)\} = \gamma_{-j}, \quad (47.4)$$

with the corresponding *autocorrelation*

$$\rho_j = \gamma_j / \sigma^2 = \rho_{-j}. \quad (47.5)$$

As noted in 45.33, the totality of coefficients $\rho_0 (=1), \rho_1, \rho_2, \dots$ is called the *correlogram* of the series. We may distinguish between the theoretical correlogram, based on the autocorrelations, and the observed correlogram, based on the serial correlations calculated for any particular series of length n .

47.6 For any given n , the theoretical correlation matrix of the set $u_t, u_{t+1}, \dots, u_{t+n}$ is the Laurent matrix

STATIONARY TIME-SERIES

$$\begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ . & . & . & . & . \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{pmatrix}. \quad (47.6)$$

Any diagonal running from North-West to South-East has the same elements.

Example 47.1

The fact that the Laurent matrix is non-negative definite implies certain consistency conditions on the autocorrelation coefficients. The determinant of any minor based on the main diagonal cannot be negative. Thus, for example,

$$\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix} = 1 - \rho_1^2 \geq 0, \text{ a trivial result.}$$

$$\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix} = 1 + 2\rho_1^2(\rho_2 - 1) - \rho_2^2 \\ = (1 - \rho_2)(1 - 2\rho_1^2 + \rho_2) \geq 0. \quad (47.7)$$

Thus, unless $\rho_2 = 1$ (and even in that case) we have

$$\rho_2 \geq 2\rho_1^2 - 1,$$

which is by no means a trivial result.

Example 47.2

As an example of a scheme which generates an autocorrelated stationary series, consider a process defined by

$$u_{t+1} = \rho u_t + \varepsilon_{t+1}, \quad (47.8)$$

where ε is a random variable with zero mean, and values $\varepsilon_p, \varepsilon_q$ are uncorrelated for $p \neq q$. We then have

$$E(u_{t+1}) = \rho E(u_t)$$

and, except perhaps in the trivial case $\rho = 1$, it follows that stationarity requires

$$E(u_t) = 0, \text{ all } t. \quad (47.9)$$

It will be seen from (47.8) that u_t depends on $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}$, etc., but not on ε_{t+1} . Thus we have

$$E\{u_t(u_{t+1} - \rho u_t)\} = E(u_t \varepsilon_{t+1}) = 0$$

and hence

$$\text{cov}(u_t, u_{t+1}) = \rho \text{var } u_t = \rho \sigma^2$$

and the correlation between

$$u_t, u_{t+1} = \rho. \quad (47.10)$$

If ρ_k is the k th order autocorrelation we have likewise

$$E\{u_t(u_{t+k} - \rho u_{t+k-1})\} = \sigma^2(\rho_k - \rho \rho_{k-1}) = 0,$$

and hence

$$\rho_k = \rho^k. \quad (47.11)$$

406 THE ADVANCED THEORY

Example 47.3

The fact that the Laurent matrix is symmetrical about its main diagonal has one somewhat unexpected consequence for schemes generated in the manner of the previous example. They could, in fact, equally well have been generated backwards, e.g. by a relation of the type

$$u_t = \rho u_{t+1} + \eta_t \quad (47.12)$$

random variable. As before, it will be seen that u_t does not depend on η_{t+1} with $t-1$ instead of t , we derive

$$\nabla u_t = \rho \nabla u_{t+1} = 0,$$

The fact that the u_t are uncorrelated in the sample, and inferences in the somewhat unexpected consequences. They could, in fact, equally well be correlated. As before, it will be seen that u_t does not depend on η_{t-1} and hence, from (47.12) with $t-1$ instead of t , we derive

$$E\{u_t(u_{t-1} - \rho u_t)\} = E(u_t \eta_{t-1}) = 0,$$

where η is another random variable. As before, we derive

$$u_t = \rho u_{t+1} + \eta_t$$

correlation ρ , as before.

$$u_t = \rho u_{t+1} + \eta_t$$

The fact that u_t is a somewhat unexpected component of y_t is an example. They could, in fact, equally well be a relation of the type

$$u_t = \rho u_{t-1} + \eta_t$$

where η is another random variable. As before, it will be seen that u_t does not depend on η_{t-1} and hence, from (47.12) with $t-1$ instead of t , we derive

$$E\{u_t(u_{t-1} - \rho u_{t-1})\} = E(u_t \eta_{t-1}) = 0,$$

correlation ρ , as before.

$$E\{u_t(u_{t-1} - \rho u_t)\} = E(u_t \eta_{t-1}) = 0,$$

giving for the first autocorrelation ρ , as before.

47.7 In time-series the transition in the reverse direction, present some new problems which

Table 47.1—Trend-free wheat-price index (European prices) compiled by the late Lord (then Sir William) Beveridge for the years 1500–1869

Year	Index	Year	Index
1500	100	1550	100
1510	105	1560	105
1520	110	1570	110
1530	115	1580	115
1540	120	1590	120
1550	125	1600	125
1560	130	1610	130
1570	135	1620	135
1580	140	1630	140
1590	145	1640	145
1600	150	1650	150
1610	155	1660	155
1620	160	1670	160
1630	165	1680	165
1640	170	1690	170
1650	175	1700	175
1660	180	1710	180
1670	185	1720	185
1680	190	1730	190
1690	195	1740	195
1700	200	1750	200
1710	205	1760	205
1720	210	1770	210
1730	215	1780	215
1740	220	1790	220
1750	225	1800	225
1760	230	1810	230
1770	235	1820	235
1780	240	1830	240
1790	245	1840	245
1800	250	1850	250
1810	255	1860	255
1820	260	1869	260

[illegible]

STATIONARY TIME-SERIES

At this point, without prejudice to that discussion, we may present a few sample series as illustrations of the type of material encountered in practice. Some of those in Chapter 45 are, on the face of it, stationary in character, e.g. Table 45.1 (barley yields) and Table 45.2 (rainfall). Table 47.1 is a famous series of trend-free wheat-price index numbers compiled by the late Lord Beveridge. It extends over 370 years, a phenomenal length of time for economic series. Table 47.2 gives deviations from a simple 11-point moving average of marriage rates in England and Wales for the period 1843–1896. Table 47.3 is an artificial series obtained by superposing a random term on a simple harmonic. Table 47.4 is another artificial series generated by a more elaborate scheme of the type of Example 47.2.

Table 47.2—Marriage rate in England and Wales: deviation from a simple 11-year moving average for the years 1843–1896

Units 1 in 10,000

Year	Marriage rate	Year	Marriage rate	Year	Marriage rate
1843	— 6	1861	— 5	1879	—12
44	1	62	— 7	80	— 5
45	12	63	1	81	0
46	10	64	6	82	5
47	— 6	65	8	83	7
48	— 8	66	9	84	3
49	— 6	67	— 2	85	— 4
50	3	68	— 8	86	— 8
51	4	69	—10	87	— 6
52	7	70	— 7	88	— 5
53	11	71	0	89	1
54	3	72	8	90	6
55	— 8	73	12	91	6
56	— 2	74	7	92	2
57	— 3	75	5	93	— 6
58	— 7	76	4	94	— 5
59	3	77	— 3	95	— 6
60	4	78	— 6	96	1

47.8 Suppose now that we have an observed series u'_1, \dots, u'_t , the primes denoting the fact that this is a single realization. Each u has mean μ and variance σ^2 . Let us define a time average

$$M_{t_2-t_1+1}(u') = \frac{1}{t_2-t_1+1} \sum_{t_1}^{t_2} u'_t \quad (47.13)$$

$$M(u') = \lim_{t_1 \rightarrow -\infty, t_2 \rightarrow \infty} M_{t_2-t_1+1}(u'). \quad (47.14)$$

Then we appeal to a theorem of Birkhoff (1931) and Khintchin (1932) which we state without proof:

- (a) If u_t is stationary with finite mean μ , $M(u')$ exists for almost all realizations, i.e. with probability unity.

(b) If and only if

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \rho_j = 0 \quad (47.15)$$

and (a) is satisfied, the average $M(u')$ is equal to the average $E(u)$, viz.

$$E(u_t) = M(u'_t). \quad (47.16)$$

This is a most important result. It implies that in practice we can estimate the mean of u from the mean of the successive values of a single realization. If this were not so, estimation from single realizations would be practically impossible.

Table 47.3—Values of the series $u_t = 10 \sin(\pi t/5) + \epsilon_t$ where ϵ_t is a rectangular random variable with range -5 to $+5$, rounded off to nearest unit

Number of term	Value of series	Number of term	Value of series	Number of term	Value of series
1	3	21	11	41	5
2	8	22	13	42	12
3	6	23	10	43	7
4	2	24	6	44	5
5	-4	25	-5	45	3
6	-7	26	-8	46	-2
7	-9	27	-12	47	-12
8	-9	28	-10	48	-12
9	-10	29	-7	49	-8
10	-1	30	0	50	-1
11	8	31	1	51	11
12	7	32	8	52	13
13	6	33	13	53	12
14	4	34	7	54	7
15	-3	35	4	55	5
16	-10	36	-9	56	-1
17	-11	37	-9	57	-6
18	-15	38	-6	58	-14
19	-4	39	-4	59	-8
20	4	40	-2	60	1

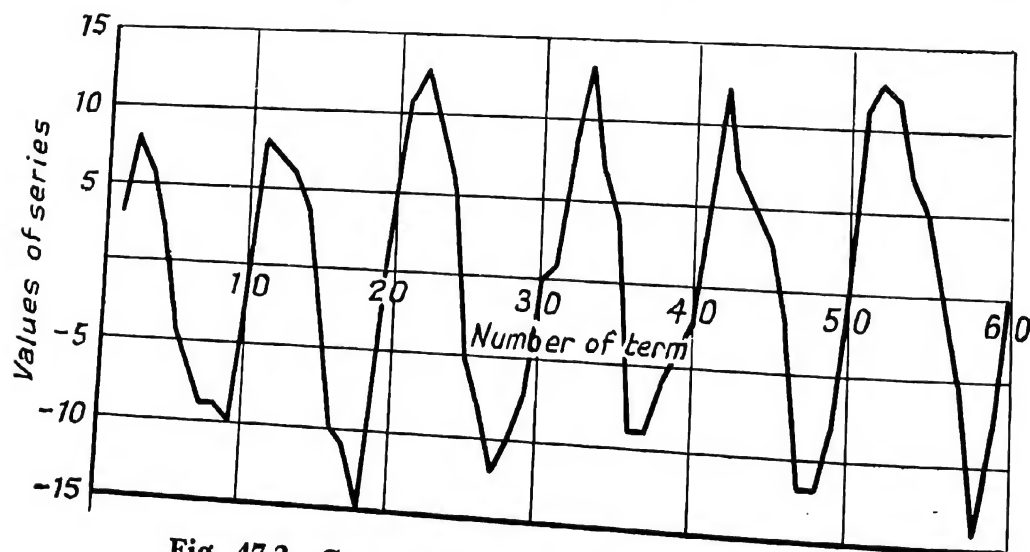


Fig. 47.2—Graph of the values of Table 47.3

Table 47.4—Values of series $u_t = 1.1u_{t-1} - 0.5u_{t-2} + \epsilon_t$ where ϵ_t is a rectangular random variable with range -9.5 to 9.5 , rounded off to nearest unit

Number of term	Value of series	Number of term	Value of series	Number of term	Value of series
1	7	23	-4	45	-13
2	6	24	-5	46	1
3	-6	25	-9	47	6
4	-4	26	-4	48	4
5	3	27	-4	49	11
6	-4	28	3	50	15
7	-5	29	9	51	9
8	-1	30	4	52	8
9	10	31	-8	53	4
10	10	32	-6	54	-1
11	6	33	-3	55	4
12	-4	34	-2	56	7
13	-4	35	0	57	11
14	-7	36	-1	58	0
15	-2	37	-3	59	1
16	6	38	3	60	0
17	17	39	-1	61	-5
18	24	40	-8	62	-11
19	17	41	-3	63	-8
20	4	42	-8	64	-3
21	1	43	-10	65	5
22	-5	44	-16		

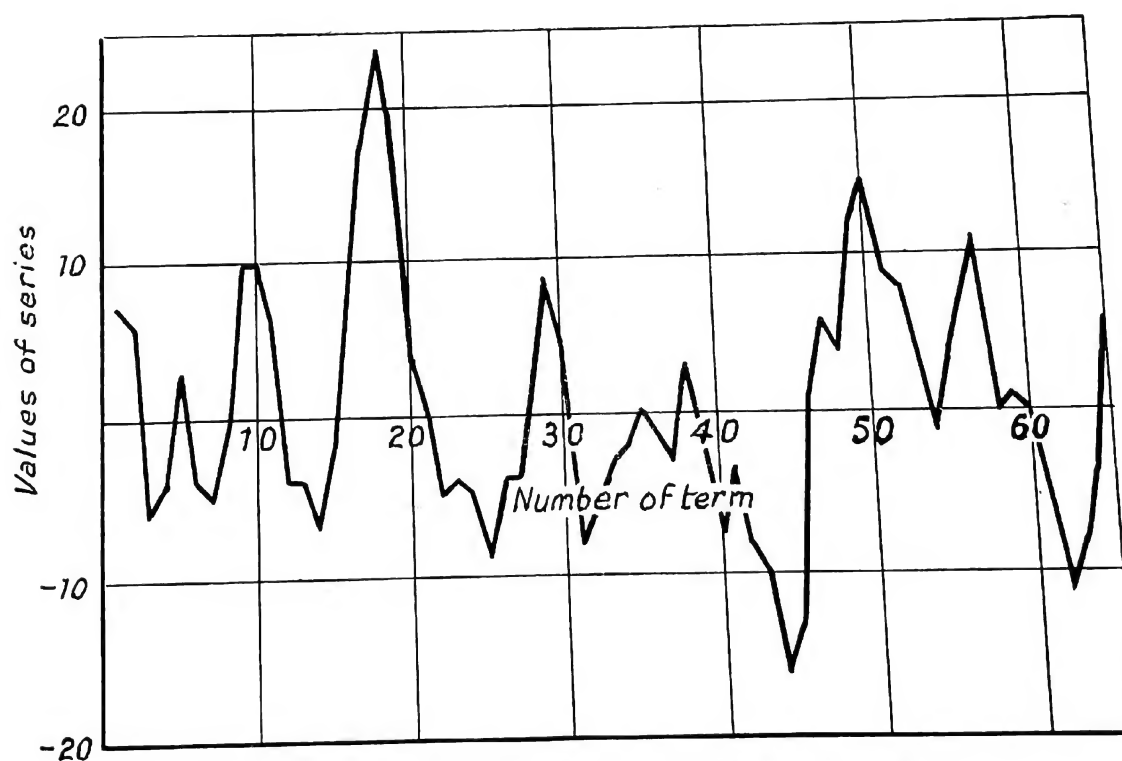


Fig. 47.3—Graph of the values of Table 47.4

47.9 If $M(u') = E(u)$ and the second moment $E(u - \mu)^2$ is finite, the process is called *ergodic*. We state, again without proof, an important extension of the Birkhoff-Khinchin result: For an ergodic process it is true, for almost all realizations, that the autocorrelation is equal to the correlation calculated from a realization

$$\frac{M\{u'_i - M(u')\} \{u'_{i+j} - M(u')\}}{M\{u'_i - M(u')\}^2}. \quad (47.17)$$

Here, again, the result enables us to estimate autocorrelations from a single realization. The condition (47.15) is not very restrictive, but it is not purely formal either. It will be obeyed if the autocorrelations dwindle to zero as the terms to which they relate become further apart, but not, for instance, if the series is a harmonic.

47.10 The correlogram, as we shall see later, is a useful instrument for exploring the nature of the internal structure of a time-series. There is a second function which serves a similar purpose and stands in relation to the correlogram in much the same relation as the characteristic function to the frequency function.

In fact, let us define a function

$$W(\alpha) = \alpha + 2 \sum_{j=1}^{\infty} \frac{\rho_j \sin \alpha j}{j}. \quad (47.18)$$

If, for some n onwards $|\rho_{n+k}| < 1$, this converges. We also write, subject to existence,

$$w(\alpha) = \frac{dW}{d\alpha} = 1 + 2 \sum_{j=1}^{\infty} \rho_j \cos \alpha j. \quad (47.19)$$

In virtue of the relation $\rho_j = \rho_{-j}$, and the fact that $\sin \theta$ is an odd function of θ , we may also write

$$w(\alpha) = \sum_{-\infty}^{\infty} \rho_j \cos \alpha j \quad (47.20)$$

$$= \sum_{-\infty}^{\infty} \rho_j e^{i\alpha j}. \quad (47.21)$$

This last form exhibits $w(\alpha)$ as a Fourier transform of the sequence ρ_j . Multiplying (47.20) by $\cos \alpha k$ and integrating term by term, we find

$$\begin{aligned} \int_0^{\pi} w(\alpha) \cos k\alpha d\alpha &= \sum_{j=-\infty}^{\infty} \rho_j \int_0^{\pi} \cos \alpha j \cos \alpha k d\alpha \\ &= \pi \rho_k \end{aligned}$$

and hence

$$\rho_j = \frac{1}{\pi} \int_0^{\pi} w(\alpha) \cos j\alpha d\alpha = \frac{1}{\pi} \int_0^{\pi} \cos \alpha j dW(\alpha). \quad (47.22)$$

We may also write

$$\rho_j = \frac{1}{\pi} \int_0^{\pi} w(\alpha) e^{-i\alpha j} d\alpha. \quad (47.23)$$

$W(\alpha)$ is called the *spectral function*. Its derivative $w(\alpha)$ is called the *spectral density*. The graph of $w(\alpha)$ as ordinate against α as abscissa is called the (*power*) *spectrum*. $w(\alpha)$ has period 2π . From (47.18) we see that $W(0) = 0$, $W(\pi) = \pi$, $W(2\pi) = 2\pi$.

47.11 The power spectrum may also be introduced directly, without reference to the correlogram, in the following manner.

For some given α , let us consider the correlation or covariance of u_t and $\cos \alpha t$. If there is some rhythm (or pseudo-rhythm—let us not beg any questions) in u_t with frequency α , the correlation will be high provided that u_t and $\cos \alpha t$ are in phase. If the series is at unit intervals and measured about its mean, consider

$$\left. \begin{aligned} a(\alpha) &= \frac{1}{\sqrt{(n\pi)}} \sum_{t=1}^n u_t \cos \alpha t \\ b(\alpha) &= \frac{1}{\sqrt{(n\pi)}} \sum_{t=1}^n u_t \sin \alpha t \end{aligned} \right\} \quad (47.24)$$

We have

$$\begin{aligned} I(\alpha) &= a^2(\alpha) + b^2(\alpha) \\ &= \frac{1}{n\pi} \{(\sum u_t \cos \alpha t)^2 + (\sum u_t \sin \alpha t)^2\} \\ &= \frac{1}{n\pi} \left\{ \sum u_t^2 + 2 \sum_{t=1}^n \sum_{k=1}^{n-t} u_t u_{t+k} \{ \cos \alpha t \cos \alpha(t+k) + \sin \alpha t \sin \alpha(t+k) \} \right\} \\ &= \frac{1}{n\pi} \left\{ \sum u_t^2 + 2 \sum_{t=1}^n \sum_{k=1}^{n-t} u_t u_{t+k} \cos k\alpha \right\} \\ &= \frac{s^2}{\pi} \left\{ 1 + 2 \sum_{k=1}^{n-t} r_k \cos k\alpha \right\} \end{aligned} \quad (47.25)$$

where $s^2 = \sum u_t^2$ and r_k is the correlation-type coefficient $\sum u_t u_{t+k} / \sum u_t^2$. In the limit this becomes

$$I(\alpha) = \frac{\sigma^2}{\pi} \left\{ 1 + 2 \sum_{k=1}^{\infty} \rho_k \cos k\alpha \right\} \quad (47.26)$$

$$\begin{aligned} &= \frac{\sigma^2}{\pi} \left\{ \sum_{k=-\infty}^{\infty} \rho_k \cos k\alpha \right\} \\ &= \frac{\sigma^2}{\pi} w(\alpha). \end{aligned} \quad (47.27)$$

The quantity I , which we call the *intensity*, is thus the spectral density multiplied by σ^2/π .

47.12 It is customary to graph the spectrum with I as ordinate against α as abscissa. In the earlier stages of the development of the subject it was more usual to compute

$$A = \frac{2}{n} \sum_{t=1}^n u_t \cos \frac{2\pi t}{\lambda}, \quad \lambda = 2\pi/\alpha, \quad (47.28)$$

$$B = \frac{2}{n} \sum_{t=1}^n u_t \sin \frac{2\pi t}{\lambda} \quad (47.29)$$

and calculate

$$S^2 = A^2 + B^2 = \frac{4}{n} \sigma^2 w(\lambda) \quad (47.30)$$

in the limit. The graph of S^2 against λ , the wavelength, was called the periodogram,

and this is a terminology we shall preserve, although some authors use "periodogram" for what we call the "spectrum."

It will be noted that whereas in (47.24) we have a divisor in \sqrt{n} , in (47.28) and (47.29) we have a divisor in n . The reason for this is that if u contains a harmonic element, say $\cos at$, of amplitude c , and the other parts of the series are uncorrelated with it, the value of S^2 at α is c^2 . In the spectrum the ordinate would be infinite at that point, at least for an infinite series. We shall return to the subject in Chapter 49.

47.13 The use which is made of the correlogram or the spectrum in exploring the internal structure of a time-series depends to some extent on the purpose of the inquiry and prior knowledge of the generating system. Broadly speaking, the correlogram is more revealing in economics, the spectrum in physics, but there are areas where the prudent research worker would use both (e.g. oceanography, meteorology, and some biological processes). The correlogram, as we have remarked, tells us something about the relationship between values of the series which are separated in time. The spectrum exhibits the extent to which the series is in step with certain fundamental rhythms; calculating the spectrum is like tuning a radio set, a signal of high power being obtained when the trial frequency coincides with an incoming frequency. For this reason the peaks of the spectrum, if any, are sometimes identified with harmonic terms in the generating system, but this is a procedure which must be carried out with some care in interpretation.

Autocorrelation generating function

47.14 In the spectral density function

$$w(\alpha) = \sum_{-\infty}^{\infty} \rho_j e^{i\alpha j} \quad (47.31)$$

put

$$z = e^{i\alpha}. \quad (47.32)$$

Then

$$w(\alpha) = \sum_{-\infty}^{\infty} \rho_j z^j = G(z), \text{ say,} \quad (47.33)$$

and we thus derive an autocorrelation generating function. We shall also find it useful to work with an autocovariance generating function

$$C(z) = \sum_{-\infty}^{\infty} \gamma_j z^j = \sigma^2 G(z). \quad (47.34)$$

Moving-average series

47.15 Consider now a series u_t and a moving average defined by

$$\zeta_t = \sum_{i=0}^{\infty} \alpha_i u_{t-i}. \quad (47.35)$$

We have here taken the moving average to be of infinite extent, so as to attain generality. But we then require to remark that ζ_t is not necessarily ergodic. It will be so, however, if $\sum \alpha_i^2$ converges, a condition which we assume to be satisfied. We then have

$$E(\zeta_t \zeta_{t+j}) = E \left\{ \sum_{i=0}^{\infty} \alpha_i u_{t-i} \sum_{k=0}^{\infty} \alpha_k u_{t+j-k} \right\}$$

$$\begin{aligned}
&= \sum \alpha_i \alpha_k E(u_{t-i} u_{t+j-k}) \\
&= \sum_{i,k=0}^{\infty} \alpha_i \alpha_k \gamma_{i+j-k} \\
&= \sum_{i=0}^{\infty} \sum_{l=0}^{\infty} \alpha_i \alpha_{i+j-l} \gamma_l.
\end{aligned} \tag{47.36}$$

If the autocovariance function of u is $C(z)$, it will be seen that the autocovariance function of ζ is given by

$$\Gamma(z) = (\sum \alpha_j z^j) (\sum \alpha_j z^{-j}) C(z). \tag{47.37}$$

In particular, if u is a purely random series $C(z) = 1$, and if we iterate a moving average k times, the autocovariance function of the resulting series is

$$(\sum \alpha_j z^j)^k (\sum \alpha_j z^{-j})^k. \tag{47.38}$$

Example 47.4

Consider a moving average of 2, $\alpha_0 = \alpha_1 = \frac{1}{2}$, of a random series.

$$\Gamma(z) = \frac{1}{4}(1+z)(1+z^{-1}) = \frac{1}{4}(z^{-1}+2+z). \tag{47.39}$$

This gives us, as is otherwise obvious,

$$\rho_1 = \frac{1}{2}, \quad \rho_j = 0, \quad j \neq 0, 1.$$

The corresponding autocorrelation generating function is derived by standardizing (47.39), so that the coefficient of z^0 is unity. Thus $G(z) = \frac{1}{2}(z^{-1}+2+z)$. Put now $z = e^{i\alpha}$. We find for the spectral density function

$$\begin{aligned}
w(\alpha) &= \frac{1}{2}(e^{-i\alpha} + 2 + e^{i\alpha}) \\
&= 1 + \cos \alpha.
\end{aligned} \tag{47.40}$$

The function is thus a cosine curve with a maximum at $\alpha = 0$. If we iterate the average k times we find

$$w(\alpha) \propto (1 + \cos \alpha)^k. \tag{47.41}$$

The constant term by which $(1 + \cos \alpha)^k$ is to be multiplied to give $w(\alpha)$ is most easily determined by the condition in 47.10 that

$$\int_0^\pi w(\alpha) d\alpha = \pi.$$

In our present case this gives

$$w(\alpha) = \frac{\pi^{\frac{1}{2}} \Gamma(k+1)}{\Gamma(k+\frac{1}{2})} \left(\frac{1 + \cos \alpha}{2} \right)^k.$$

This, for increasing k , tends to unity at $\alpha=0$ and zero elsewhere. All the ρ_j tend to unity. The series thus tends to a constant value, as is otherwise evident from the fact that successive iterations smooth out fluctuation.

On the other hand, if we take successive *differences* of the original random series, $\alpha_0 = \frac{1}{2}$, $\alpha_1 = -\frac{1}{2}$, and we find after k differences

$$w(\alpha) \propto \left(\frac{1 - \cos \alpha}{2} \right)^k. \tag{47.42}$$

This tends to unity at $\alpha = \pi$. The even order autocorrelations tend to $+1$, the odd order to -1 . The series thus tends to terms which are equal in absolute value but alternate in sign.

Example 47.5

The moving average with weights

$$\frac{1}{6}[-1, 2, 4, 2, -1]$$

is repeatedly applied to a random series. The autocovariance generating function after k iterations is then

$$6^{-k}\{-z^{-2} + 2z^{-1} + 4 + 2z - z^2\}^k.$$

Putting $z = e^{i\alpha}$ we find

$$w(\alpha) = c_0(3 + 2 \cos \alpha + 2 \cos^2 \alpha)^k \quad (47.43)$$

where c_0 is some constant. For variations in α the maximum value of $w(\alpha)$ occurs when $\cos \alpha = \frac{1}{2}$ and we may write

$$w(\alpha) = c_0\{1 - \frac{4}{9}(\cos \alpha - \frac{1}{2})^2\}^k. \quad (47.44)$$

Thus, for $\cos \alpha - \frac{1}{2} = \varepsilon$, say, the ordinate of $w(\alpha)$ tends to zero, as k increases, compared with the ordinate at $\cos \alpha = \frac{1}{2}$. For continual iteration, therefore, the resultant series tends to a periodic wave with period $2\pi/\arccos \frac{1}{2} = 6$.

Example 47.6 Slutsky's theorem of the sinusoidal limit

Take a moving average of two of a random series n times, and take the m th difference of the result. Then, if $n \rightarrow \infty$ such that m/n tends to some constant θ between 0 and 1, the series tends to a sine wave with wavelength λ given by $\lambda = \arccos \frac{1-\theta}{1+\theta}$.

Taking the m th difference is equivalent to taking first differences m times. Hence the autocovariance generating function of the resultant is given by

$$\Gamma(z) \propto (1+z^{-1})^n (1+z)^n (1-z^{-1})^m (1-z)^m,$$

and hence, putting $z = e^{i\alpha}$, we find

$$w(\alpha) \propto (1 - \cos \alpha)^m (1 + \cos \alpha)^n. \quad (47.45)$$

We can evaluate the constant from the relation

$$\int_0^\pi w(\alpha) d\alpha = \pi$$

and find

$$w(\alpha) = \frac{n\Gamma(m+n+1)}{2^{m+n}\Gamma(m+\frac{1}{2})\Gamma(n+\frac{1}{2})} (1 - \cos \alpha)^m (1 + \cos \alpha)^n. \quad (47.46)$$

The maximum value occurs at $\alpha = \alpha_0$, when

$$\cos \alpha_0 = \frac{n-m}{n+m} = \frac{1-\theta}{1+\theta}. \quad (47.47)$$

The theorem then follows if we show that $w(\alpha) \rightarrow 0$ everywhere except at this maximum; and further that the series is not only periodic but a sine wave.

Using Stirling's approximation to the Γ function, we find

$$w(\alpha) \sim \frac{(\frac{1}{2}\pi)^{\frac{1}{2}}(m+n)^{\frac{1}{2}}}{2^{m+n}e^{(m-\frac{1}{2})m}(n-\frac{1}{2})^n} (1 - \cos \alpha)^m (1 + \cos \alpha)^n.$$

If $\cos \alpha = \frac{n-m}{n+m} + \varepsilon$ this tends to

$$\frac{(\frac{1}{2}\pi)^{\frac{1}{2}}}{e} (m+n)^{\frac{1}{2}} \left(1 - \frac{n+m}{2m} \varepsilon\right)^m \left(1 + \frac{n+m}{2n} \varepsilon\right)^n. \quad (47.48)$$

For $\varepsilon = 0$ this tends to infinity. For $\varepsilon \neq 0$ it will be seen, on taking logarithms and expanding, that the expression tends to zero uniformly in any closed interval excluding $\varepsilon = 0$. Thus $w(\alpha)$ has a single infinite ordinate at α_0 given by

$$\arccos \{(1-0)(1+0)\}.$$

$W(\alpha)$ is accordingly a step function with a single step from 0 to π at that point.

It follows from (47.22) that the autocorrelations of the resulting series are given by

$$\rho_j = \cos \alpha_0 j. \quad (47.49)$$

Consider now a given stretch of the derived series, say u_1, \dots, u_N for fixed N as $n \rightarrow \infty$. We have

$$E \sum_{i=1}^{N-2} (u_{i+2} - 2\rho u_{i+1} + u_i)^2 = 2(N-2) \text{var } u \{1 - 2\rho_1^2 + \rho_2\},$$

which, in virtue of (47.49), becomes in the limit

$$2(N-2) \text{var } u \{1 - 2 \cos^2 \alpha_0 + \cos 2\alpha_0\} = 0.$$

Hence in the limit

$$u_{i+2} - 2\rho_1 u_{i+1} + u_i = 0. \quad (47.50)$$

This is a difference equation of a sine curve.

For some generalizations of Slutsky's result see Romanovsky (1932, 1933) and Moran (1949).

47.16 If u_t is a stationary series the moving average

$$\zeta_t = \beta_0 u_t + \beta_1 u_{t-1} + \dots + \beta_h u_{t-h} \quad (47.51)$$

is also stationary. In particular, if u_t is a purely random process with zero mean the autocorrelations are given by

$$\rho_j = \frac{\sum_{i=0}^{h-j} \beta_i \beta_{i+j}}{\sum_{i=0}^h \beta_i^2}, \quad j \leq h, \\ = 0, \quad j > h. \quad (47.52)$$

We have already seen in Example 46.7 that the correlogram may present an oscillatory appearance for such an average.

47.17 Wold (1938) has proved a theorem on the conditions under which a specified set of constants $\rho_1, \rho_2, \dots, \rho_h$ can be the autocorrelations of a moving average of a random series. Take the generating function

$$G(z) = 1 + \rho(z + z^{-1}) + \dots + \rho_h(z^h + z^{-h}). \quad (47.53)$$

Put

$$y = z + z^{-1}. \quad (47.54)$$

This will transform $G(z)$ into a polynomial of degree h in y , say $H(y)$. Then, for the ρ 's to be autocorrelations of a moving average of extent $h+1$ it is necessary and sufficient that $H(y)$ has no real root of odd multiplicity in the interval $-2 < y < 2$.

For example, suppose $\rho_1 \neq 0$ and all other ρ 's vanish.
 $G(z)$ is then $1 + \rho_1(z + z^{-1})$ and

$$H(y) = 1 + \rho_1 y.$$

This has a root of odd-order multiplicity (unity) equal to $-1/\rho_1$. This will lie in the interval -2 to $+2$ unless $\rho_1 < \frac{1}{2}$. Thus, no moving average of extent 2 can have $\rho_1 > \frac{1}{2}$, $\rho_2 = \rho_3 = \dots = 0$, as is otherwise evident.

We have to determine the β 's in the moving average from the relation

$$G(z) = \sum_{j=0}^h \rho_j (z^j + z^{-j}) = \sum_{j=0}^h \beta_j z^j \sum_{j=0}^h \beta_j z^{-j}. \quad (47.55)$$

There will, in general, be 2^h solutions (for, on identifying powers of z , we shall have h equations each of which is quadratic). However, only one of these gives roots of $G(z) = 0$ which lie inside the unit circle, and in virtue of another result to which we refer later (47.18), this is the only acceptable solution. From (47.54) it is seen that for any y , the roots in z are given by

$$z^2 - zy + 1 = 0 \quad (47.56)$$

and hence, having the product unity, lie one inside and one outside the unit circle.

From (47.56) we have

$$z = \frac{1}{2}y \pm (\frac{1}{4}y^2 - 1)^{\frac{1}{2}}. \quad (47.57)$$

Three cases arise:

- $H(y)$ has a complex root y_1 . Then the conjugate y_1^* is also a root. Thus the corresponding quantities $z_1, z_1^{-1}, z_1^*, (z_1^*)^{-1}$ are roots of $G(z)$ and thus one of $(z - z_1)(z - z_1^*)$ and $(z - z_1^{-1})(z - z_1^{*-1})$ is a factor of $\sum \beta_j z^j$.
 - $H(y)$ has a real root ≥ 2 in modulus. Then, from (47.57), z_1 and z_1^{-1} are both real. One must be a root of $\sum \beta_j z^j = 0$ and this case then corresponds to real roots of $\sum \beta_j z^j = 0$.
 - $H(y)$ has a real root < 2 in modulus. In this case z_1 and z_1^{-1} are conjugate complex and of modulus unity. The factors $z - z_1$ and $z - z_1^{-1}$ are therefore both contained in $\sum \beta_j z^j$ and $\sum \beta_j z^{-j}$ and therefore the root must be of even multiplicity.
- The theorem follows.

Autoregressive series

47.18 Consider now a series defined by

$$u_t = -\alpha_1 u_{t-1} - \alpha_2 u_{t-2} \dots - \alpha_h u_{t-h} + \varepsilon_t \quad (47.58)$$

which (putting $\alpha_0 = 1$) we may write in the more convenient form

$$\sum_{j=0}^h \alpha_j u_{t-j} = \varepsilon_t. \quad (47.59)$$

Here ε_t is a random variable and, unless otherwise specified, we shall suppose that successive values of ε are independent and all have the same variance.

If D is an operator such that $Du_t = u_{t-1}$, we may write (47.59) as

$$(\sum \alpha_j D^j)u_t = \varepsilon_t,$$

giving the formal solution

$$\begin{aligned} u_t &= \frac{1}{\sum \alpha_j D^j} \varepsilon_t = (\sum \beta_j D^j) \varepsilon_t \\ &= \sum_{j=0}^{\infty} \beta_j \varepsilon_{t-j}, \end{aligned} \quad (47.60)$$

where the constants β are related to the α 's by the identity in D

$$\frac{1}{\sum \alpha_j D^j} = \sum \beta_j D^j. \quad (47.61)$$

However, this is not the complete solution of the difference equation (47.59). Let z_1, z_2, \dots, z_h be the roots of

$$z^h + \alpha_1 z^{h-1} + \dots + \alpha_h = 0. \quad (47.62)$$

Then the general solution of

$$\sum \alpha_j u_{t-j} = 0 \quad (47.63)$$

may be written

$$u_t = \sum_{j=1}^h A_j z_j^t \quad (47.64)$$

where the A 's are arbitrary constants.

We shall now assume that $|z_j| < 1$ for any j , namely that the roots of (47.62) all fall within the unit circle, and that they are all different. Then for large t the solution (47.64) damps out of existence.

The series (47.59) is regarded as having "started up" a long time in the past. Then the contribution to the solution (47.64) has disappeared and the complete solution is, in fact, the particular solution (47.60).

We shall call a series of form (47.58) *autoregressive*. It is a type of moving average of infinite extent. If ε is ergodic, then so will be u_t , provided that $\sum \beta_j^2$ converges. This proviso is, in fact, satisfied, provided that the roots of (47.62) are all within the unit circle (cf. Exercise 47.19).

In practice it is rarely necessary to discuss the roots of equation (47.62). J. Wise (1956), however, has shown by using a theorem of Routh, that the conditions concerning the roots can be expressed as algebraic conditions on the α 's themselves.

47.19 It will have been observed that u_t is dependent on $\varepsilon_t, \varepsilon_{t-1}$, etc., but not on $\varepsilon_{t+1}, \varepsilon_{t+2}$, etc. Let us multiply

$$\sum \alpha_j u_{t-j} = \varepsilon_t \quad (47.65)$$

by u_{t-k} , and take expectations. We then find

$$\rho_k + \alpha_1 \rho_{k-1} + \dots + \alpha_h \rho_{k-h} = 0, \quad k > 0, \quad (47.66)$$

a set of equations due to Yule (1927) and G. Walker (1931). In particular, since $\rho_{-j} = \rho_j$, we have

$$\rho_1 + \alpha_1 + \alpha_2 \rho_1 + \dots + \alpha_h \rho_{h-1} = 0$$

$$\rho_2 + \alpha_1 \rho_1 + \alpha_2 + \dots + \alpha_h \rho_{h-2} = 0$$

and so on.

If we multiply (47.64) by u_{t+k} , $k \geq 0$, and take expectations, we obtain

$$\rho_k + \alpha_1 \rho_{k+1} + \dots + \alpha_h \rho_{k+h} = \frac{\text{var } \varepsilon}{\text{var } u} \beta_k \quad (47.67)$$

where the β 's are given by (47.61). These equations are due to Wold (1938).

Example 47.7 The Markoff series

Consider again the series of Example 47.2. This is the simplest case of an autoregressive series (apart from the trivial case $h = 0$):

$$u_t + \alpha_1 u_{t-1} = \varepsilon_t,$$

which, for convenience, we shall write as

$$u_t - \rho u_{t-1} = \varepsilon_t. \quad (47.68)$$

This is known as a Markoff series.

We have

$$\frac{1}{1 - \rho D} = 1 + \rho D + \rho D^2 + \dots$$

and hence

$$\beta_j = \rho^j. \quad (47.69)$$

Hence

$$\begin{aligned} \text{var } u &= \text{var} \left\{ \sum_{j=0}^{\infty} \rho^j \varepsilon_{t-j} \right\} \\ &= \text{var } \varepsilon \sum_{j=0}^{\infty} \rho^{2j} \\ &= \frac{\text{var } \varepsilon}{1 - \rho^2}. \end{aligned} \quad (47.70)$$

From the Yule-Walker equations (47.66) with $h = 1$ we have

$$\rho_j - \rho \rho_{j-1} = 0, \quad j > 0,$$

and in particular

$$\rho_1 = \rho. \quad (47.71)$$

(This is, in fact, why we named as ρ the parameter $-\alpha_1$.) Further,

$$\rho_k = \rho^k. \quad (47.72)$$

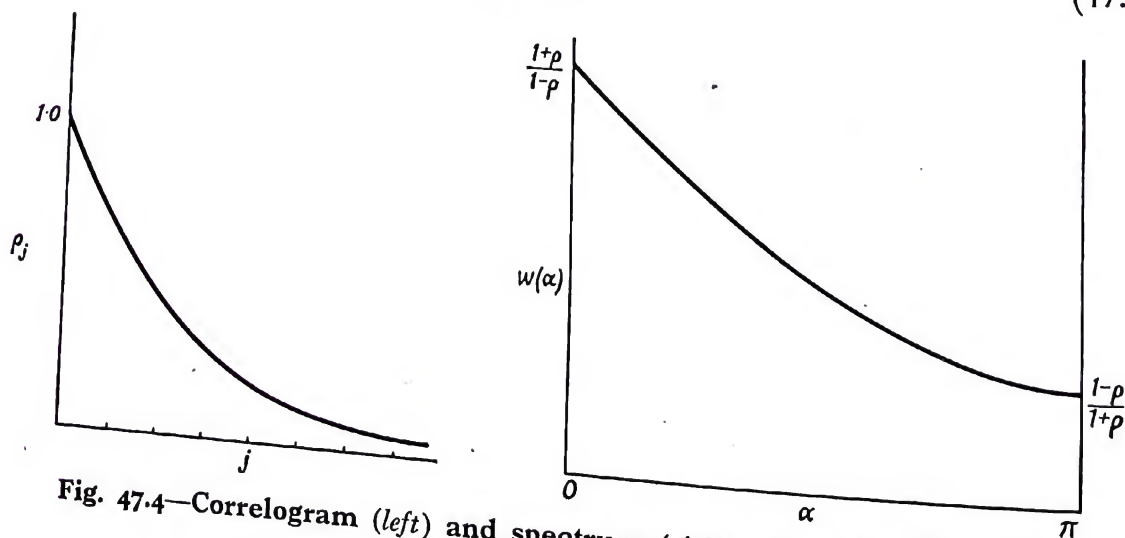


Fig. 47.4—Correlogram (left) and spectrum (right) of the Markoff series

For the spectral density we have

$$\begin{aligned} w(\alpha) &= \sum_{-\infty}^{\infty} \rho^{|j|} z^j = -1 + \frac{1}{1-\rho z} + \frac{1}{1-\rho z^{-1}} \\ &= -1 + \frac{1}{1-\rho e^{i\alpha}} + \frac{1}{1-\rho e^{-i\alpha}} \\ &= \frac{1-\rho^2}{1-2\rho \cos \alpha + \rho^2}. \end{aligned} \quad (47.73)$$

The correlogram and the power spectrum are shown in Fig. 47.4.

Example 47.8 The Yule series

The next most complex form of linear autoregressive series is known by the name of Yule and is given by

$$u_t + \alpha_1 u_{t-1} + \alpha_2 u_{t-2} = \varepsilon_t. \quad (47.74)$$

From the first two Yule-Walker equations (47.66) we have

$$\rho_1 + \alpha_1 + \alpha_2 \rho_1 = 0$$

$$\rho_2 + \alpha_1 \rho_1 + \alpha_2 = 0,$$

giving

$$\rho_1 = -\frac{\alpha_1}{1+\alpha_2} \quad (47.75)$$

$$\rho_2 = -\alpha_2 + \frac{\alpha_1^2}{1+\alpha_2} \quad (47.76)$$

or

$$\alpha_1 = -\frac{\rho_1(1-\rho_2)}{1-\rho_1^2} \quad (47.77)$$

$$\alpha_2 = -1 + \frac{1-\rho_2}{1-\rho_1^2} = -\frac{\rho_2 - \rho_1^2}{1-\rho_1^2}. \quad (47.78)$$

These equations give the parameters α_1, α_2 in terms of the first two autocorrelations and vice versa. More generally, if μ, ν are the roots of

$$x^2 + \alpha_1 x + \alpha_2 = 0$$

then

$$\rho_j = A\mu^j + B\nu^j,$$

subject to initial conditions

$$\rho_0 = 1 = A + B$$

$$\rho_1 = A\mu + B\nu.$$

We then find

$$\rho_j = \frac{1}{(\mu - \nu)(1 + \mu\nu)} [\mu^{j+1}(1 - \nu^2) - \nu^{j+1}(1 - \mu^2)]. \quad (47.79)$$

We can put this in a slightly more convenient form. Put

$$\mu = pe^{i\theta}, \quad \nu = pe^{-i\theta}. \quad (47.80)$$

We find

$$p = |\sqrt{\alpha_2}|, \quad \cos \theta = \frac{-\alpha_1}{2|\sqrt{\alpha_2}|}, \quad (47.81)$$

and (47.79) reduces to

$$\rho_j = \frac{p^j \sin(j\theta + \psi)}{\sin \psi} \quad (47.82)$$

where

$$\tan \psi = \frac{1+p^2}{1-p^2} \tan \theta. \quad (47.83)$$

The spectral density function is given by

$$w(\alpha) = \frac{(1-\alpha_2)(1-\alpha_1^2+\alpha_2^2+2\alpha_2)}{(1+\alpha_2)\{1+\alpha_1^2+\alpha_2^2-2\alpha_2+2\alpha_1(1+\alpha_2)\cos \alpha + 4\alpha_2 \cos^2 \alpha\}}. \quad (47.84)$$

Fig. 47.5 shows a typical correlogram and power spectrum.

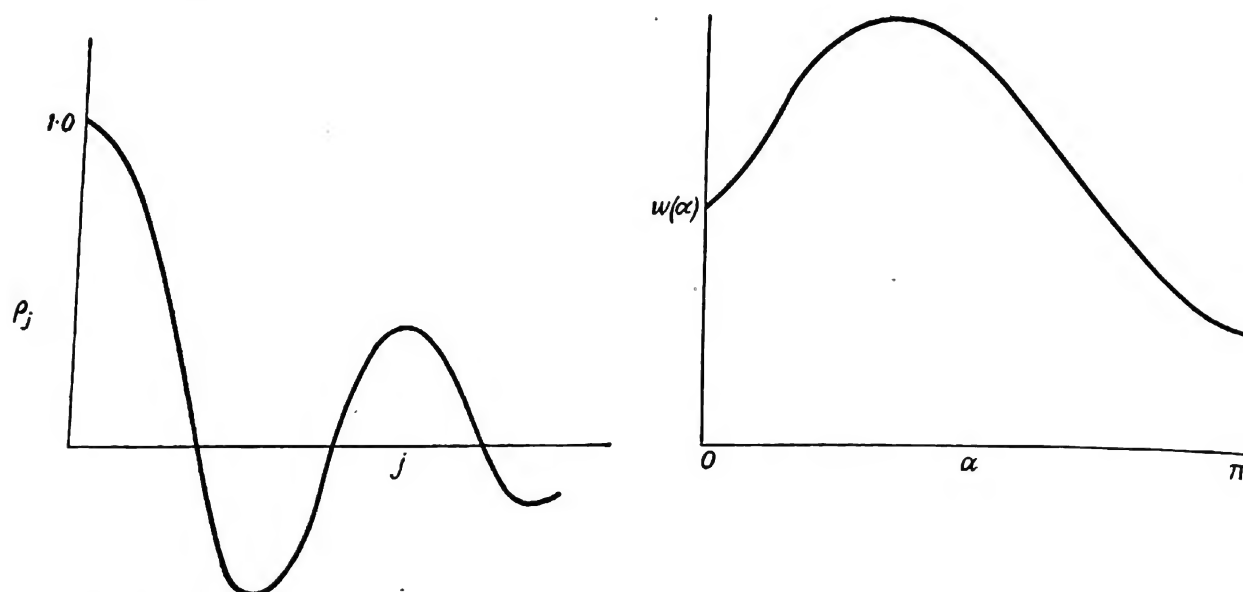


Fig. 47.5—Correlogram (left) and spectrum (right) of the Yule series

Example 47.9

We return to the discussion at the end of 47.18. In a linear scheme of the autoregressive type, if the roots of the characteristic equation do not lie inside the unit circle, the process is not ergodic. If some lie outside the circle the series “explodes.” It may still oscillate, but with ever-increasing amplitude.

If a root lies *on* the unit circle, the general solutions do not damp away, but provide harmonic terms. The particular solution is then a “wandering series.” Consider, for example, the simple case $u_t = u_{t-1} + \varepsilon_t$. Clearly the particular solution is

$$u_t = \sum_{j=0}^{\infty} \varepsilon_{t-j}$$

and the variance of u increases without limit.

Example 47.10

Consider the Yule scheme in the limiting case $\alpha_2 = 1$. The coefficient p of (47.81) then is unity and the correlogram (47.82) does not damp. The system, in fact, ceases to be ergodic.

47.20 The type of series exemplified by

$$u_t = \zeta_t + \varepsilon_t, \quad (47.85)$$

where ζ_t is a deterministic harmonic term, may be regarded as a harmonic with superposed error. It is sometimes known as a scheme of hidden periodicities.

There is a somewhat different type of process to which the name harmonic is sometimes given, though it is of no practical importance. A series may consist of the sum of a number of harmonics, say

$$u_t = \sum_j A_j \cos(\alpha_j t), \quad (47.86)$$

where the α 's are fixed but the A 's may vary from one realization to another. In this case there will be certain linear relations of the Yule-Walker type

$$\sum_{j=0}^h c_j \rho_{j-h} = 0. \quad (47.87)$$

Continuous series

47.21 Up to this point we have been concerned with series which are defined or observed at a set of discrete points. Some series, as we noted in Chapter 45, have a continuous existence in time, and there are even situations where we can form a continuous record, as for example in the devices which graph temperature on a rotating drum. The fact that matter is ultimately discontinuous (if it is a fact) does not prevent us from regarding this record as continuous.

For series which are defined by deterministic continuous functions, such as polynomials or trigonometric functions, this correspondence between the assumed continuity of reality and the defined continuity of mathematics rarely causes any conceptual difficulty. But when we come to series of the stationary type in which there are jumps between successive points, expressed by random variables, we must consider this question of continuity more closely. Can we, in fact, have a continuous series which proceeds by random jumps, however small? Our own opinion is that we cannot; that there is something essentially antithetic between randomness and continuity. Any tendency, then, to take the mathematician's customary leap from the discontinuous to the continuous case must be carefully controlled. It may well prove possible, of course, to approximate to discontinuous expressions by continuous ones, for example, to represent sums by integrals; but we must not forget the problems of interpretation which are involved.

47.22 To deal with this subject rigorously requires a theory of stochastic integration which would take us beyond the scope of this book. But we may expound the basic results in an intuitive way as follows.

Consider a continuous series $u(t)$ defined in some interval $-h$ to h . Taking the mean to be zero, which does not seriously limit our generality, we may define the variance as

$$\text{var } u = \frac{1}{2h} \int_{-h}^h u^2(t) dt. \quad (47.88)$$

If this has a limiting value as h tends to infinity (as it will for a stationary series) the variance is defined over an infinite range. Likewise we have an autocovariance function, and if we standardize by division by $\text{var } u$, we have the *autocorrelation function*

$$\rho(k) = \lim_{h \rightarrow \infty} \frac{1}{2h} \int_{-h}^h u(t)u(t+k) dt. \quad (47.89)$$

Consider the transform of the autocorrelation function, say $\phi_r(p)$, defined by

$$\begin{aligned} \phi_r(p) &= \frac{1}{2h} \int_{-h}^h \rho(k) e^{ikp} dk = \lim_{h \rightarrow \infty} \frac{1}{4h^2} \int_{-h}^h \int_{-h}^h u(t)u(t+k) dt e^{ikp} dk \\ &= \lim_{h \rightarrow \infty} \frac{1}{4h^2} \int_{-h}^h \int_{-h}^h u(t) e^{-ipt} u(t+k) e^{ip(t+k)} dt dk. \end{aligned} \quad (47.90)$$

Putting $q = t+k$, we reduce this to

$$\begin{aligned} &\lim_{h \rightarrow \infty} \frac{1}{4h^2} \int_{-h}^h \int_{-h}^h u(t) e^{-ipt} u(q) e^{ipq} dt dq \\ &= \lim_{h \rightarrow \infty} \frac{1}{4h^2} \int_{-h}^h u(t) e^{-ipt} dt \int_{-h}^h u(q) e^{ipq} dq. \end{aligned} \quad (47.91)$$

Hence, if the transform of the series $u(t)$ is given by

$$\phi_u(p) = \lim_{h \rightarrow \infty} \frac{1}{2h} \int_{-h}^h u(t) e^{ipt} dt = a(p) + ib(p), \quad (47.92)$$

we have, on letting h tend to infinity in (47.89), for the transform of the autocorrelation function

$$\phi_r(p) = a^2(p) + b^2(p) = |\phi_u(p)|^2. \quad (47.93)$$

$\phi_r(p)$ is the continuous extension of the spectral density which (cf. (47.21)) is the Fourier transform of the autocorrelations.

47.23 It is to be noted that, even for a continuous function defined over an infinite interval, the autocorrelation function does not determine the series $u(t)$ uniquely. In fact, given $\phi_r(p)$, we have from (47.93)

$$\phi_u(p) = |\phi_r(p)|^{\frac{1}{2}} e^{i\mu}, \quad (47.94)$$

where μ is any arbitrary real function. We shall then have, on inverting for u ,

$$\begin{aligned} u(t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_u(p) e^{-itp} dp \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} |\phi_r(p)|^{\frac{1}{2}} e^{i\mu - itp} dp \end{aligned} \quad (47.95)$$

Since $u(t)$ must be real, the imaginary part of the integral vanishes and we have

$$u(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \sqrt{\phi_r} \cos(\mu - tp) dp, \quad (47.96)$$

a result due to Wiener (1930). Hence μ must be an odd function of p but is otherwise (subject to convergence) arbitrary. Hence ϕ_r does not uniquely determine $u(t)$. We shall consider this from the point of view of spectral functions later.

STATIONARY TIME-SERIES

Example 47.11

Consider an autocorrelation function of the type we have discussed for the Yule series at (47.82),

$$\rho(k) = \frac{p^k \sin(k\theta + \psi)}{\sin \psi}.$$

Consider what happens if we regard this as defined for all k , not merely integral values. We may, with a slight change of notation, put

$$\rho(k) = \frac{e^{-qk} \sin(k\theta + \psi)}{\sin \psi}, \quad q > 0. \quad (47.97)$$

When k is negative we must use $|k|$ in this expression.

For the transform we have

$$\begin{aligned} \phi_r(p) &= \int_{-\infty}^{\infty} \frac{e^{-|qk|} \sin(k\theta + \psi)}{\sin \psi} e^{ikp} dk \\ &= \frac{q}{q^2 + (\theta + p)^2} + \frac{q}{q^2 + (\theta - p)^2}. \end{aligned} \quad (47.98)$$

The variable p in the transform here is not to be confused with the damping factor p in $\rho(k)$.

It is to be noted in (47.98) that this spectrum is continuous with a maximum at $p = 0$. The physicist would be tempted to regard this spectrum as analogous to that of white light, every frequency being represented. It does not follow, however, that the series $u(t)$ arises as the sum of a large number of harmonic terms with all possible frequencies, in the way that white light can be regarded as the composition of a number of resonators oscillating on all wavelengths.

On this question of white light, let us consider the limiting case when a time-series is defined at a series of small intervals Δt and all autocorrelations are zero. We then find, from (47.21), that the spectral density (on a scale with unit time-intervals) is unity, or on a scale Δt would be approximately $\Delta t/2\pi$, namely a constant. Certain physical systems do give rise to constant spectral densities, or to a series of equal ordinates very close together. The communications engineer describes the situation as one of *white noise*. Since he is trying to transmit signals on a determined frequency this so-called noise is a nuisance (like that in a radio set) which affects his reception and acts as an error-like disturbance to the purity of the incoming signals.

Filters and transfer functions

47.24 Suppose that we have a series $u(t)$ and a system of weights $a(t)$. We may form what is, in effect, a linear weighted average $v(t)$ by the formula

$$v(t) = \int_0^{\infty} a(\tau) u(t - \tau) d\tau. \quad (47.99)$$

This average is over past values of $u(t)$, including the present value, and does not look to the future. If $u(t)$ is defined at discontinuous points a similar sum may be defined. For the spectrum, we consider

$$\int_{-\infty}^{\infty} v(t) e^{it\alpha} dt = \int_0^{\infty} a(\tau) \int_{-\infty}^{\infty} e^{it\alpha} u(t - \tau) dt d\tau$$

$$\begin{aligned}
 &= \int_0^\infty a(\tau) e^{i\tau\alpha} \int_{-\infty}^\infty u(t) e^{it\alpha} dt d\tau \\
 &= \int_0^\infty a(\tau) e^{i\tau\alpha} d\tau \int_{-\infty}^\infty u(t) e^{it\alpha} dt.
 \end{aligned}$$

Hence, taking the squares of the moduli,

$$|\phi_v(\alpha)|^2 = |\phi_u(\alpha)|^2 \left| \int_0^\infty e^{i\tau\alpha} a(\tau) d\tau \right|^2. \quad (47.100)$$

The function

$$T(\alpha) = \int_0^\infty e^{i\tau\alpha} a(\tau) d\tau \quad (47.101)$$

is sometimes called the frequency response function or *transfer function*. It is, in essence, the c.f. (Fourier transform) of the weighting function $a(\tau)$. (47.93) and (47.100) show that the spectral density of the derived series is obtained from that of the original series on multiplication by the square of the modulus of the transfer function. The engineer would regard the system as an incoming series, $u(t)$, modified by some mechanism equivalent to a linear average, to give the output $v(t)$.

47.25 Within limits, it is possible to choose the transfer function so as to produce from $u(t)$ an output $v(t)$ with emphasis on particular frequencies. Such a function, or rather the set of weights $a(\tau)$, is then called a *filter*. This is not the happiest expression, a filter removing impurities by withholding them, rather than transforming them; but it will serve. We need not, and shall not, confine our usage to averages which extend over the past, as in (47.99). Thus, the ordinary moving averages which we considered in Chapter 46 are filters in this sense.

Partial autocorrelations

47.26 If we write the linear autoregressive scheme in the form (47.58)

$$u_t = -\alpha_1 u_{t-1} - \alpha_2 u_{t-2} \dots - \alpha_h u_{t-h} + \varepsilon_t \quad (47.102)$$

we may regard it as a kind of predictive equation for u_t , which will then depend on two factors, the systematic terms in u_{t-j} which, as it were, express the effect on u_t of its own past history, and the random element ε , which can be considered as a disturbance. We may then ask the questions which are usually posed in ordinary regression analysis: given the autocorrelations ρ_j , what are the partial correlations expressing the dependence of u_t on previous terms when the effect of other intermediate previous terms has been removed?

47.27 Consider first of all the Markoff scheme (47.68),

$$u_t = \rho u_{t-1} + \varepsilon_t. \quad (47.103)$$

We know from (47.72) that $\rho_h = \rho^h$. Let us calculate the partial autocorrelation $\rho_{13.2}$, expressing the dependence of u_t on u_{t-2} apart from the intervention of u_{t-1} . We have, in an obvious notation (cf. (27.5)),

$$\rho_{13.2} = \frac{\rho_{13} - \rho_{12}\rho_{32}}{(1 - \rho_{12}^2)^{\frac{1}{2}}(1 - \rho_{32}^2)^{\frac{1}{2}}},$$

with

$$\rho_{13} = \rho^2, \quad \rho_{12} = \rho_{32} = \rho.$$

Thus

$$\rho_{13.2} = 0. \quad (47.104)$$

It will easily be seen that, in fact, all the partial autocorrelations are zero. This is otherwise obvious from the fact that in (47.103) the regression of u_t concerns only u_{t-1} as independent variable.

In short, for a Markoff scheme all partial autocorrelations vanish. The term u_t , so far as it depends systematically on previous terms, is entirely explained by u_{t-1} .

47.28 Now consider the Yule scheme (47.74),

$$u_t = -\alpha_1 u_{t-1} - \alpha_2 u_{t-2} + \varepsilon_t. \quad (47.105)$$

We have from (47.75) and (47.76)

$$\rho_1 = -\frac{\alpha_1}{1 + \alpha_2}$$

$$\rho_2 = -\alpha_2 + \frac{\alpha_1^2}{1 + \alpha_2}.$$

The partial correlation between u_t and u_{t-2} is given by

$$\rho_{13.2} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} = -\alpha_2, \quad (47.106)$$

as we might expect.

We can easily check that higher-order partials vanish. For example, the numerator of $\rho_{14.23}$ is $\rho_{14.3} - \rho_{12.3}\rho_{42.3}$. This in turn has a numerator

$$(1 - \rho_1^2)(\rho_3 - \rho_2\rho_1) - (\rho_1 - \rho_2\rho_1)(\rho_2 - \rho_1^2). \quad (47.107)$$

Considering the determinant of the first three Yule-Walker equations (47.66), we have

$$\begin{vmatrix} \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \\ \rho_3 & \rho_2 & \rho_1 \end{vmatrix} = 0.$$

This, expanded by the first column, shows that the expression (47.107) vanishes.

Such results are, of course, obvious from the general theory of regression when we recall that the regressor variables and the regressand all have the same variance, so that the coefficients in the regression equation, being standardized, are equal to the partial correlations.

Infinite, semi-infinite and circular processes

47.29 If we consider the linear autoregressive scheme

$$\sum_{j=0}^h \alpha_j u_{t-j} = \varepsilon_t \quad (47.108)$$

as generating a series of values of u , given those of ε , we are faced with a difficulty, or rather, with the necessity for making a decision. We cannot find the value of u at some point, say T , without knowing those for $T-1, T-2, \dots, T-h$. We may suppose that these values are given, or otherwise known, for some T_0 . From that point onwards the series is ascertainable and we may say that it is semi-infinite, because it is considered as extending to infinity in one direction, that of increasing t .

On the other hand, we may regard the series as extending back into the infinite past, as well as forward into the infinite future. We do not then know the "starting up" values, if any. The series may be said to be infinite.

47.30 For certain mathematical reasons which will become clear, we may also define a circular *process* on lines analogous to those we have used in defining circular coefficients (45.34). We suppose, in fact, that for some N

$$\begin{aligned}u_{t+N} &= u_t \\u_{t+N+1} &= u_{t+1} \\&\vdots \\u_{t+N+h} &= u_{t+h}.\end{aligned}$$

It is, so far as we can see, impossible to imagine physical processes which generate such a system. We shall therefore avoid it as far as we can. The best that can be said about it is that, should we be able to derive results for the circular process in an exact form, there is some expectation that, by letting N tend to infinity, we may derive at least approximations to the results for the semi-infinite case. But even this is doubtful; one does not avoid a difficulty by banishing it to infinity. We shall therefore need to be very careful in interpreting results for the circular process.

47.31 The theoretical forms of correlograms exemplified in Figs. 47.4 and 47.5 are not followed very closely by the observed correlograms of short series ("short" in this context meaning anything up to 100 terms). Exercises 47.20–22 give the observed correlations of Tables 47.1, 47.2 and 47.4. Apart from irregularities such as might be expected from sampling effects, there are two other phenomena encountered in practice: (a) the serial correlations are biased downwards, and (b) the correlations of higher order do not damp out for Yule and Markoff schemes as quickly as might be expected. A theoretical explanation of these effects will be given in the following chapter. The problem of how to fit schemes of various kinds to stationary series and how to test hypotheses concerning them will be considered in Chapter 50.

EXERCISES

47.1 In a stationary series with $\rho_1 = \rho_2 = \rho$, show that $\rho \geq -\frac{1}{2}$ and that

$$\rho_3 \geq \frac{4\rho^2 - \rho - 1}{\rho + 1}.$$

47.2 For the Markoff series

$$u_t = \rho u_{t-1} + \varepsilon_t,$$

show that the cumulants of u are connected with those of ε by

$$\kappa_r(u) = \kappa_r(\varepsilon)/(1 - \rho^r).$$

Hence show that for the standardized coefficients $\lambda_r = \kappa_r/\kappa_2^{\frac{1}{2}r}$,

$$\lambda_r(u) = \lambda_r(\varepsilon) \frac{(1 - \rho^2)^{\frac{1}{2}r}}{1 - \rho^r}.$$

Deduce that in general u is closer to normality than ε , but that it is not normal unless ε is normal.

47.3 Show that the Markoff scheme of the previous exercise can be written

$$u_t = \rho u_{t+1} + \eta_t$$

with

$$\eta_t = -\rho \varepsilon_{t+1} + (1 - \rho^2) \sum_{j=0}^{\infty} \rho^j \varepsilon_{t-j}.$$

Hence show that if η is a random variable,

$$\text{var } \eta = \text{var } \varepsilon$$

$$\text{cov}(\eta_t, \eta_{t+k}) = 0, \quad k \neq 0.$$

47.4 Verify that for the spectral density function (47.73)

$$\int_0^\pi w(\alpha) d\alpha = \pi.$$

47.5 Verify that for the Yule autoregressive scheme equation (47.84) is true and that the integral of $w(\alpha)$ over the range 0 to π is equal to π .

47.6 Show that there exist four and only four moving averages of a random series with correlograms

$$\rho_1 = -\frac{42}{85}, \quad \rho_2 = \frac{4}{17}, \quad \rho_3 = -\frac{8}{85}, \quad \rho_4 = \rho_5 = \text{etc.} = 0.$$

These are $\frac{1}{5}[8, -4, 2, -1]$, $\frac{1}{5}[-1, 2, -4, 8]$, $\frac{1}{5}[2, -1, 8, -4]$, $\frac{1}{5}[-4, 8, -1, 2]$.

(Wold, 1938)

47.7 In the general autoregressive scheme with random term ε , show that

$$\text{var} \left(\sum_{j=0}^h \alpha_j u_{t+j} \right) = \text{var } \varepsilon.$$

47.8 For the Yule autoregressive scheme show that

$$\frac{\text{var } u}{\text{var } \varepsilon} = \frac{1 + \alpha_2}{(1 - \alpha_2) \{(1 + \alpha_2)^2 - \alpha_1^2\}}.$$

47.9 Show that the autocorrelations of the m th differences of a random series are given by

$$\rho_j = (-1)^j \frac{m^{(j)}}{(m+j)^{(j)}}.$$

47.10 If any series is fitted by a Yule scheme with autocorrelations ρ_j show that the autocorrelations of the residuals, say σ_j , are given by

$$\sigma_j = \frac{(1 + \alpha_1^2 + \alpha_2^2)\rho_j + \alpha_1(1 + \alpha_2)(\rho_{j+1} + \rho_{j-1}) + \alpha_2(\rho_{j+2} + \rho_{j-2})}{1 + \alpha_1^2 + \alpha_2^2 + 2\alpha_1(1 + \alpha_2)\rho_1 + 2\alpha_2\rho_2}.$$

47.11 Show that a series for which the autocorrelation function is

$$v(j) = (\sin \lambda j) / \lambda j$$

has a continuous spectrum with a jump at the point λ .

47.12 Show that any linear autoregressive series can be represented as a combined sequence of Yule and Markoff series in which the error term ε of one is the series-value of the next.

47.13 If A, B are defined as

$$A = \frac{2}{n} \sum_{j=1}^n u_j \cos \beta_j$$

$$B = \frac{2}{n} \sum_{j=1}^n u_j \sin \beta_j$$

show that, if $u_t = a \sin \alpha t + b_t$, where b_t is a component uncorrelated with $\sin \alpha t$,

$$A = \frac{a}{n} \left[\frac{\sin \left\{ \frac{1}{2} n (\alpha - \beta) \right\} \sin \left\{ \frac{1}{2} (n+1) (\alpha - \beta) \right\}}{\sin \left\{ \frac{1}{2} (\alpha - \beta) \right\}} + \frac{\sin \left\{ \frac{1}{2} n (\alpha + \beta) \right\} \sin \left\{ \frac{1}{2} (n+1) (\alpha + \beta) \right\}}{\sin \left\{ \frac{1}{2} (\alpha + \beta) \right\}} \right]$$

with a similar expression for B . Hence show that $S^2 = A^2 + B^2$ remains small as n increases unless $\alpha - \beta$ is small, in which case $S^2 = a^2$.

47.14 For a "continuous" series with autocovariance function

$$\rho(k) = e^{-a|k|}$$

show that the spectral density is given by

$$w(\alpha) = \frac{1}{\pi} \frac{1}{a^2 + \alpha^2}.$$

(Cf. the characteristic function of a Cauchy distribution.)

47.15 A series obeys the relation

$$u_t = u_{t-1} + \varepsilon_t$$

where ε_t is a random series with unit variance. It is divided into consecutive groups of m terms and the arithmetic mean of each group determined, say as v_t . Show that

$$\text{var}(\Delta v_t) = (2m^2 + 1)/3m$$

and that

$$\text{cov}(\Delta v_t, \Delta v_{t+1}) = \frac{m^2 - 1}{2(2m^2 + 1)}.$$

(Working, 1960)

47.16 Let U be the $N \times N$ matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \dots & \cdot \\ 0 & 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}.$$

Show that successive powers of U are of a similar form with the diagonal of unities displaced to the right, and that $U^N = 0$. Hence show that if \mathbf{u} is a column vector $(u_t, u_{t-1}, \dots, u_N)$ the autoregressive scheme may be written

$$\left(\sum_{j=0}^h \alpha_j U^j \right) \mathbf{u} = \boldsymbol{\epsilon}.$$

Further show that for the dispersion matrix of \mathbf{u}

$$\mathbf{V}(\mathbf{u}) = (\sum \alpha_j U^j)^{-1} (\sum \alpha_j U'^j)^{-1}.$$

(Whittle, 1951)

47.17 Show further that the inverse of the dispersion matrix of U is given by

$$\mathbf{V}^{-1} = (\sum \alpha_j U'^j) (\sum \alpha_j U^j)$$

and hence that for a third-order linear autoregressive series with random errors the inverse of the autodispersion matrix is given by

$$V^{-1} = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & \alpha_3 & . & . & . & . \\ \alpha_1 & 1 + \alpha_1^2 & \alpha_1(1 + \alpha_2) & \alpha_2 + \alpha_1 \alpha_3 & . & . & . & . \\ \alpha_2 & \alpha_1(1 + \alpha_2) & 1 + \alpha_1^2 + \alpha_2^2 & \alpha_1 + \alpha_1 \alpha_2 + \alpha_2 \alpha_3 & . & . & . & . \\ \alpha_3 & \alpha_2 + \alpha_1 \alpha_3 & \alpha_1 + \alpha_1 \alpha_2 + \alpha_2 \alpha_3 & 1 + \alpha_1^2 + \alpha_2^2 + \alpha_3^2 & . & . & . & . \\ \text{etc.} & & & & \text{etc.} & & & \end{pmatrix}. \quad (\text{J. Wise, 1955})$$

47.18 Show that although the autocorrelation matrix of a series is of the Laurent type, its inverse is not. (Whittle, 1951)

47.19 Referring to equation (47.63), show, by expanding the left-hand side in terms of partial fractions, that $\sum \beta_j^2$ converges if the roots of (47.62) are all different and lie within the unit circle.

47.20 The following are the serial correlations of the data of Table 47.1 (wheat prices). Draw the correlogram.

Order of correlation k	r_k	k	r_k	k	r_k	k	r_k
1	0.562	16	0.158	31	0.060	46	-0.036
2	0.103	17	0.109	32	-0.008	47	-0.013
3	-0.075	18	0.002	33	-0.039	48	0.042
4	-0.092	19	-0.075	34	0.007	49	0.062
5	-0.082	20	-0.062	35	0.056	50	0.065
6	-0.136	21	-0.021	36	0.010	51	0.050
7	-0.211	22	-0.062	37	-0.004	52	0.009
8	-0.261	23	-0.088	38	-0.015	53	-0.027
9	-0.192	24	-0.084	39	-0.047	54	-0.053
10	-0.070	25	-0.076	40	-0.047	55	-0.073
11	-0.003	26	-0.091	41	0.008	56	-0.106
12	-0.015	27	-0.052	42	0.034	57	-0.084
13	-0.012	28	-0.032	43	0.065	58	-0.019
14	0.047	29	-0.012	44	0.099	59	0.003
15	0.101	30	0.059	45	0.009	60	0.010

47.21 The following are the serial correlations of Table 47.2 (marriage rates). Draw the correlogram.

Order of correlation k	r_k	k	r_k
1	0.563	11	-0.080
2	-0.089	12	-0.136
3	-0.498	13	-0.132
4	-0.631	14	-0.058
5	-0.467	15	-0.095
6	-0.025	16	-0.126
7	0.353	17	-0.036
8	0.396	18	0.131
9	0.254	19	0.209
10	0.104	20	0.205

THE ADVANCED THEORY OF STATISTICS

430

47.22 The following are the serial correlations of Table 47.4 (artificial series). Draw the correlogram.

Order of correlation k	r_k	k	r_k	k	r_k
1	0.70	11	-0.05	21	0.05
2	0.29	12	-0.17	22	-0.12
3	0.01	13	-0.27	23	-0.28
4	-0.17	14	-0.31	24	-0.43
5	-0.27	15	-0.30	25	-0.57
6	-0.25	16	-0.18	26	-0.56
7	-0.13	17	0.12	27	-0.26
8	0.07	18	0.29	28	0.02
9	0.12	19	0.33	29	0.17
10	0.05	20	0.22	30	0.27

CHAPTER 48

THE SAMPLING THEORY OF SERIAL CORRELATIONS

Large-sample theory

48.1 We defined the serial correlation of lag k in 45.32 and remarked that, for certain purposes, simpler forms of definition were mathematically and computationally more convenient. For large n the definitions tend to equivalence. For large sample theory we shall therefore consider the standard error of the form

$$r_j = \frac{1}{n} \sum_{i=1}^{n-j} u_i u_{i+j} \bigg/ \frac{1}{n} \sum_{i=1}^n u_i^2$$

$$= c/v, \text{ say.} \quad (48.1)$$

As usual, we may write parental or sample forms indifferently in the resulting expressions, and shall usually employ the autocorrelations ρ_j .

In accordance with the customary procedure we have

$$\delta r_j = \frac{\delta c}{v} - \frac{c \delta v}{v^2}$$

$$\text{var } r_j = \frac{\text{var } c}{v^2} - \frac{2c \text{cov}(c, v)}{v^3} + \frac{c^2 \text{var } v}{v^4}$$

and, taking $v = 1$ without loss of generality,

$$\text{var } r_j = \text{var } c - 2c \text{cov}(c, v) + c^2 \text{var } v. \quad (48.2)$$

To evaluate this expression we will derive a general result concerning the covariance of two covariance terms. We have

$$\text{cov} \left\{ \frac{1}{n} \sum u_a u_{a+s}, \frac{1}{n} \sum u_b u_{b+s+t} \right\} = \frac{1}{n^2} E \{ \sum u_a u_{a+s} \sum u_b u_{b+s+t} \} - \rho_s \rho_{s+t}$$

$$= \frac{1}{n^2} E \left\{ \sum_{a,b} (u_a u_{a+s} u_b u_{b+s+t}) \right\} - \rho_s \rho_{s+t}. \quad (48.3)$$

To evaluate the product-moments of order four in this expression we need some further assumptions. Assume that the u 's are jointly normally distributed so that their characteristic function is of the form

$$\exp \left\{ -\frac{1}{2} (\theta_a^2 + \theta_{a+s}^2 + \theta_b^2 + \theta_{b+s+t}^2 + 2\rho_s \theta_a \theta_{a+s} + \text{etc.}) \right\}. \quad (48.4)$$

For the coefficient of $\theta_a \theta_{a+s} \theta_b \theta_{b+s+t}$ we find

$$\rho_s \rho_{s+t} + \rho_{b-a} \rho_{b-a+t} + \rho_{b-a+s+t} \rho_{b-a-s}.$$

On summing over a, b , we find, for the covariance (48.3),

$$\frac{1}{n^2} \left[n^2 \rho_s \rho_{s+t} + n \sum_{i=-\infty}^{\infty} \rho_i \rho_{i+t} + n \sum_{i=-\infty}^{\infty} \rho_{i+s+t} \rho_{i-s} \right] - \rho_s \rho_{s+t}$$

$$= \frac{1}{n} \{ \sum \rho_i \rho_{i+t} + \sum \rho_{i+s+t} \rho_{i-s} \}. \quad (48.5)$$

Now let us specialize. Putting $s = t = 0$, we have

$$\text{var } v = \frac{2}{n} \sum_{i=-\infty}^{\infty} \rho_i^2. \quad (48.6)$$

Putting $t = 0$, we have

$$\text{var } c = \frac{1}{n} \sum_{i=-\infty}^{\infty} (\rho_i^2 + \rho_{i+s} \rho_{i-s}). \quad (48.7)$$

Putting $s = 0$ and replacing t by s , we have

$$\text{cov}(v, c) = \frac{2}{n} \sum_{i=-\infty}^{\infty} \rho_i \rho_{i+s}. \quad (48.8)$$

Finally, substitution of these values in (48.2) gives us

$$\text{var } r_j = \frac{1}{n} \sum_{i=-\infty}^{\infty} \left[\rho_i^2 + \rho_{i-j} \rho_{i+j} - 4\rho_j \rho_i \rho_{i+j} + 2\rho_i^2 \rho_j^2 \right]. \quad (48.9)$$

The formula is due to Bartlett (1946).

48.2 This result shows us that, even for large samples with the simplifying assumption of normality, the variance of r_j depends on all the autocorrelations of the series. This is awkward, for we cannot estimate them all directly from a finite series. Some fair approximations can, however, be derived in the manner of the following examples.

Example 48.1

Consider in the first place the simple case when all parent autocorrelations are zero (a random series). We then find, from (48.9),

$$\text{var } r_j = \frac{1}{n}. \quad (48.10)$$

This verifies that the variance is of order $1/n$. In fact, the sampling formulae in this case reduce to those of an ordinary correlation coefficient in bivariate normal samples, as is evident from the fact that the series is random.

Example 48.2

If ρ_j and subsequent ρ 's are small, (48.9) reduces to approximately

$$\text{var } r_j = \frac{1}{n} \sum_{i=-(j-i)}^{j-i} \rho_i^2. \quad (48.11)$$

It may be verified (we leave this as Exercise 48.1) that on similar assumptions

$$\text{cov}(r_j, r_{j+l}) = \frac{1}{n} \sum_i \rho_i \rho_{i+l}. \quad (48.12)$$

Example 48.3

If the series obeys the Markoff equation (47.68) with parameter ρ , we have from (48.11) and (47.72) for large j

$$\begin{aligned} \text{var } r_j &= \frac{1}{n} \left\{ \sum_{i=-\infty}^{\infty} \rho^{2i} \right\} \\ &= \frac{1}{n} \frac{1+\rho^2}{1-\rho^2}. \end{aligned} \quad (48.13)$$

For more exact values for small j , see Exercise 48.8.

Example 48.4

The approximate forms of (48.11) and (48.12) can be derived direct from the autocorrelation generating function. For, in the notation of 47.14,

$$w(\lambda) = G(z) = \sum_{-\infty}^{\infty} \rho_i z^i,$$

and hence

$$\{G(z)\}^2 = \sum_{\alpha=-\infty}^{\infty} z^{\alpha} \sum_i \rho_i \rho_{i-\alpha} = \sum_i z^{\alpha} (\sum_i \rho_i \rho_{\alpha+i}), \quad (48.14)$$

so that G^2 is a generating function for the sums required.

For example, with the Markoff process

$$G(z) = -1 + \frac{1}{1-\rho z} + \frac{1}{1-\rho z^{-1}}$$

$$G^2(z) = 1 + (1-\rho z)^{-2} + (1-\rho z^{-1})^{-2} - 2(1-\rho z)^{-1} - 2(1-\rho z^{-1})^{-1} + 2(1-\rho z)(1-\rho z^{-1}).$$

We then find for the coefficient of z^0

$$1 + 1 + 1 - 2 - 2 + 2(1 + \rho^2 + \rho^4 + \dots) = \frac{1 + \rho^2}{1 - \rho^2},$$

as at (48.13). Also the coefficient of z^k is given by

$$(k+1)\rho^k - 2\rho^k + 2(\rho^k + \rho^{k+2} + \dots) = \rho^k \left\{ k - 1 + \frac{2}{1 - \rho^2} \right\}.$$

Hence the covariance of r_j and r_{j+k} in a Markoff scheme is, by (48.12),

$$\frac{1}{n} \rho^k \left\{ k - 1 + \frac{2}{1 - \rho^2} \right\} \quad (48.15)$$

and the correlation between them is, using (48.13),

$$\frac{\rho^k \{(k+1) + (k-1)\rho^2\}}{1 + \rho^2}. \quad (48.16)$$

The method may be extended to schemes of higher order (cf. Quenouille (1947a) and Exercise 48.14).

Bias in the estimation of autocorrelations

48.3 If r is of the form A/\sqrt{BC} , we have, writing a, b, c for deviations from means,

$$r_j = \frac{E(A) + a}{[\{E(B) + b\} \{E(C) + c\}]^{\frac{1}{2}}}$$

and expanding in binomial series to the second order of approximation, we find

$$E(r_j) = \frac{E(A)}{\{E(B)E(C)\}^{\frac{1}{2}}} \left\{ 1 - \frac{E(ab)}{2E(A)E(B)} - \frac{E(ac)}{2E(A)E(C)} + \frac{E(bc)}{4E(B)E(C)} + \frac{3E(b^2)}{8E^2(B)} + \frac{3E(c^2)}{8E^2(C)} + \dots \right\}. \quad (48.17)$$

Putting

$$B = \frac{1}{n-j} \sum_{i=1}^{n-j} u_i^2 - \frac{1}{(n-j)^2} \left(\sum_{i=1}^{n-j} u_i \right)^2 \quad (48.18)$$

$$C = \frac{1}{n-j} \sum_{i=1}^{n-j} u_{i+j}^2 - \frac{1}{(n-j)^2} \left(\sum_{i=1}^{n-j} u_{i+j} \right)^2 \quad (48.19)$$

so that $E(B) = E(C)$, we find, on a little reduction, using the asymptotic equivalence of B and C ,

$$E(r_j) \doteq \frac{E(A)}{E(B)} - \frac{\text{cov}(a, b)}{E^2(B)} + \frac{E(A) \text{var } b}{E^3(B)}. \quad (48.20)$$

Let the variance of the series be unity and write $\nu = n - i$. Then we have, on taking expectations in (48.18-19),

$$E(B) = E(C) = \frac{1}{\nu} \left\{ \nu - 1 - \frac{2}{\nu} \sum_{i=1}^{\nu-1} (\nu - i) \rho_i \right\}. \quad (48.21)$$

Now, taking

$$A = \frac{1}{n-j} \sum_{i=1}^{n-j} u_i u_{i+j} - \frac{1}{(n-j)^2} \sum_{i=1}^{n-j} u_i \sum_{i=1}^{n-j} u_{i+j} \quad (48.22)$$

we find

$$E(A) = \frac{1}{\nu} \left\{ \nu \rho_j - \frac{1}{\nu} \sum_{i=0}^{\nu-1} (\nu - j) \rho_{j+i} - \frac{1}{\nu} \sum_{i=0}^j (\nu - i) \rho_{j-i} - \frac{1}{\nu} \sum_{i=1}^{\nu-j-1} (\nu - j - i) \rho_i \right\}, \quad j < 0. \quad (48.23)$$

We have evaluated in (48.6) and (48.8) $\text{var } b$ and $\text{cov}(a, b)$. Substituting for the various quantities in (48.20) we find $E(r_j)$. We shall not write this out explicitly, but shall consider some particular cases.

Example 48.5

If the series is random, $\rho_j = 0$ except for $\rho_0 = 1$. We then find from (48.20)

$$E(r_j) \doteq -\frac{1}{\nu}. \quad (48.24)$$

It so happens that in this case we can evaluate the bias exactly by using the definition

$$r_j = \frac{n}{n-j} \frac{\sum_{i=1}^{n-j} (u_i - \bar{u})(u_{i+j} - \bar{u})}{\sum_{i=1}^n (u_i - \bar{u})^2}.$$

Put

$$z = u_i - \bar{u}.$$

Then

$$\begin{aligned} E(r_j) &= \frac{n}{n-j} E \frac{\sum z_i z_{i+j}}{\sum z_i^2} \\ &= n E \frac{z_i z_{i+j}}{\sum z_i^2} \\ &= \frac{n}{n(n-1)} E \frac{\sum_{i,k=1}^n z_i z_k}{\sum z_i^2}, \quad i \neq k, \\ &= \frac{1}{n-1} E \frac{(\sum z_i)^2 - \sum z_i^2}{\sum z_i^2} \\ &= -\frac{1}{n-1}, \end{aligned}$$

since $\sum z_i \equiv 0$. This agrees with (48.24) to order n^{-1} . It is rather remarkable that there is a downward bias in r_j even for a random series.

Example 48.6

If the series is such that

$$\rho_1 = \rho, \quad \rho_j = 0, \quad j \neq 0, 1,$$

we find from (48.20)

$$E(r_1) = \rho + \frac{1}{v}(1 + \rho)(4\rho^2 - 2\rho - 1) \quad (48.26)$$

$$E(r_2) = -\frac{1}{v}(1 + 2\rho + 2\rho^2) \quad (48.27)$$

$$E(r_j) = -\frac{1}{v}(1 + 2\rho), \quad j < 2. \quad (48.28)$$

Example 48.7

For the Markoff series (47.68) with parameter ρ we find similarly

$$E(r_1) = \rho - \frac{1 + 3\rho}{v} \quad (48.29)$$

$$E(r_j) = \rho^j - \frac{1}{v} \left\{ \frac{1 + \rho}{1 - \rho} (1 - \rho^j) + 2k\rho^j \right\}, \quad j > 1. \quad (48.30)$$

The bias in all these cases is downwards and obviously may be quite serious. For $\rho = \frac{1}{2}$ in a Markoff series of 25 terms the mean value of r_1 would be about 0.4, not 0.5.

Quenouille's correction

48.4 In the manner of (40.28) we may use the simplification of Quenouille's method of removing bias by splitting the series into two. If r is the serial coefficient for the whole series, $r_{(1)}$ and $r_{(2)}$ those for the two halves, we use

$$R = 2r - \frac{1}{2}(r_{(1)} + r_{(2)}), \quad (48.31)$$

which will be unbiased to order n^{-1} .

For some further results on bias see Marriott and Pope (1954), Kendall (1954), and Quenouille (1956) (cf. 17.10, Vol. 2). White (1961) has obtained some results for the Markoff case to order n^{-3} .

Some exact results

48.5 We now proceed to consider some exact results in the distribution theory of serial correlations. As we might expect, exactitude has to be purchased at a price, usually that of assuming normality in the parent series, but occasionally, also, of simplifying the definition of the statistic under investigation. We first of all derive some results (due to Moran) by the method of expectations. We then obtain some distributions (due to R. L. Anderson and a series of later writers, notably Daniels) which raise some quite new points in distribution theory.

48.6 Consider the case of parent normality with zero autocorrelations and first serial correlation defined by

$$r_1 = \frac{n}{n-1} \frac{\sum (u_i - \bar{u})(u_{i+1} - \bar{u})}{\sum (u_i - \bar{u})^2}. \quad (48.32)$$

We have already shown at (48.25) that

$$E(r_1) = -\frac{1}{n-1}. \quad (48.33)$$

Put

$$I = \frac{n-1}{n} r_1, \quad z_i = u_i - \bar{u},$$

so that

$$E(I) = -\frac{1}{n}.$$

We have

$$\begin{aligned} E(I^2) &= E \left(\frac{\sum_{i=1}^{n-1} z_i z_{i+1}}{\sum z_i^2} \right)^2 \\ &= E \left[(\sum z_i^2)^{-2} \left\{ \sum_1^{n-1} z_i^2 z_{i+1}^2 + \sum_1^{n-2} z_i z_{i+1}^2 z_{i+2} + \sum z_i z_{i+1} z_k z_{k+1} \right\} \right] \end{aligned} \quad (48.34)$$

where $i, i+1, k, k+1$ are all distinct. Thus

$$E(I^2) = E[(\sum z_i^2)^{-2} \{ (n-1)z_1^2 z_2^2 + 2(n-2)z_1^2 z_2 z_3 + (n-2)(n-3)z_1 z_2 z_3 z_4 \}]$$

or, in terms of the augmented symmetric functions (12.5, Vol. 1),

$$E(I^2) = E \left[[2]^{-2} \left\{ \frac{1}{n} [2^2] + \frac{2}{n(n-1)} [21^2] + \frac{1}{n(n-1)} [1^4] \right\} \right]. \quad (48.35)$$

Using Appendix Table 10, we express the augmented symmetric functions in terms of power-sums, obtaining, since $(1) \equiv 0$ here,

$$E(I^2) = E \left[\frac{1}{n} \left\{ 1 - \frac{(4)}{(2)^2} \right\} + \frac{2}{n(n-1)} \left\{ \frac{2(4)}{(2)^2} - 1 \right\} + \frac{1}{n(n-1)} \left\{ 3 - \frac{6(4)}{(2)^2} \right\} \right]. \quad (48.36)$$

Now in normal samples k_4/k_2^2 is independent of k_2^2 (cf. 37.27) and, using (12.28),

$$\frac{k_4}{k_2^2} = \frac{n-1}{(n-2)(n-3)} \left\{ n(n+1) \frac{(4)}{(2)^2} - 3(n-1) \right\}. \quad (48.37)$$

Hence, since the sample is normal, the expectation of k_4/k_2^2 vanishes, giving, from (48.37),

$$E \left\{ \frac{(4)}{(2)^2} \right\} = \frac{3(n-1)}{n(n+1)}. \quad (48.38)$$

Substitution in (48.36) and reduction then gives

$$E(I^2) = \frac{n^2 - 3n + 3}{n^2(n-1)}, \quad (48.39)$$

and hence

$$\begin{aligned} \text{var } I &= \frac{(n-2)^2}{n^2(n-1)} \\ \text{var } r_1 &= \frac{(n-2)^2}{(n-1)^3}. \end{aligned} \quad (48.40)$$

It may be noted that, to order n^{-1} ,

$$\text{var } r_1 = \frac{1}{n+1}. \quad (48.41)$$

Moran (1947-8) gives results for the circular definition and also derives the third and fourth moments—cf. Exercises 48.6, 48.12. Jenkins (1954-6) has used similar methods for the joint distribution of serial correlations (cf. 48.26 and Exercises 48.18 and 48.19).

R. L. Anderson's distribution

48.7 For reasons which will become evident later, we now consider the distribution of the first circular serial correlation

$$r_1 = \frac{u_1 u_2 + u_2 u_3 + \dots + u_{n-1} u_n + u_n u_1 - n \bar{u}^2}{\sum_{i=1}^n (u_i - \bar{u})^2}. \quad (48.42)$$

Following R. L. Anderson (1942) we consider the distribution of this statistic in samples from an independent normal series. We now drop the suffix to r .

We shall seek a linear transformation to variables ξ_i ($i = 1, 2, \dots, n$) such that r transforms to $\sum \lambda_i \xi_i^2 / \sum \xi_i^2$. The point about using a circular definition is that we can determine the λ 's explicitly.

Any orthogonal transformation will transform the denominator of r to the required form. The numerator of r is equal to q , say, with

$$q = \sum u_i u_{i+1} \left(1 - \frac{2}{n}\right) - \frac{1}{n} \sum u_i^2 - \frac{2}{n} \sum u_i u_{i+k}, \quad k > 1. \quad (48.43)$$

Consider

$$q - \lambda \sum u_i^2. \quad (48.44)$$

As in the case of principal components (cf. 43.6), if we determine λ so as to maximize this quantity, we shall arrive at the sum $\sum \lambda_i \xi_i^2$, as desired. We do not need to find the transformation: the λ 's are all that we require, and from (48.43-4) they are the roots of

$$\begin{vmatrix} -\left(\lambda + \frac{1}{n}\right) & \frac{1}{2}\left(1 - \frac{2}{n}\right) & -\frac{1}{n} & \dots & -\frac{1}{n} & \frac{1}{2}\left(1 - \frac{2}{n}\right) \\ \frac{1}{2}\left(1 - \frac{2}{n}\right) & -\left(\lambda + \frac{1}{n}\right) & \frac{1}{2}\left(1 - \frac{2}{n}\right) & \dots & -\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & \frac{1}{2}\left(1 - \frac{2}{n}\right) & -\left(\lambda + \frac{1}{n}\right) & \dots & -\frac{1}{n} & -\frac{1}{n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{2}\left(1 - \frac{2}{n}\right) & -\frac{1}{n} & -\frac{1}{n} & \dots & \frac{1}{2}\left(1 - \frac{2}{n}\right) & -\left(\lambda + \frac{1}{n}\right) \end{vmatrix} = 0. \quad (48.45)$$

This is a circulant determinant and can be factorized. Consider, in fact, the circulant

$$D = \begin{vmatrix} a_1 & a_2 & \dots & a_n \\ a_n & a_1 & \dots & a_{n-1} \\ a_{n-1} & a_n & \dots & a_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ a_2 & a_3 & \dots & a_1 \end{vmatrix} \quad (48.46)$$

Let $\omega_1, \omega_2, \dots, \omega_{n-1}, \omega_n (= 1)$ be the n roots of unity. Multiply the j th column of D by ω_k^{j-1} and sum for j . Then it will be seen that $\sum_j a_j \omega_k^{j-1}$ is a factor of D . This is true for any k , and hence

$$D = \prod_{k=1}^n \sum_{j=1}^n a_j \omega_k^{j-1}. \quad (48.47)$$

Now

$$\sum_{j=1}^n \omega_k^{j-1} = 0, \quad k \neq n, \\ = n, \quad k = n,$$

and thus

$$\sum_{j=3}^{n-1} \omega_k^{j-1} = n-3, \quad k = n, \\ = -(1 + \omega_k + \omega_k^{-1}), \quad k \neq n. \quad (48.48)$$

Putting the appropriate values of the a 's, from (48.45), in (48.47), and using (48.48) we find, on some reduction,

$$D = \lambda \prod_{k=1}^{n-1} \left\{ -\lambda + \frac{1}{2}(\omega_k + \omega_k^{-1}) \right\},$$

and, since $\frac{1}{2}(\omega_k + \omega_k^{-1}) = \cos \frac{2\pi k}{n}$, we have

$$D = \lambda \prod_{k=1}^{n-1} \left(-\lambda + \cos \frac{2\pi k}{n} \right). \quad (48.49)$$

Thus the roots are $\lambda = 0$, $\lambda = \cos \frac{2\pi k}{n}$, $k = 1, 2, \dots, n-1$. It is crucial to observe that the λ -roots occur in pairs, except perhaps for one which is unity. These paired terms may be put together to give v_j (= sum of squares of two ξ^2). Thus we have

$$q = \sum_{i=1}^{\frac{1}{2}(n-1)} \lambda_i v_i, \quad n \text{ odd}, \quad (48.50)$$

$$= \sum_{i=1}^{\frac{1}{2}(n-2)} \lambda_i v_i - v, \quad n \text{ even}, \quad (48.51)$$

where v_0 is distributed as χ^2 with 2 d.fr. and v with one d.fr. The denominator of r , say p , is distributed as χ^2 with $n-1$ d.fr. Our problem is then reduced to finding the distribution of

$$r = \frac{\sum_{i=1}^{\frac{1}{2}(n-1)} \lambda_i v_i / \sum v_i, \quad n \text{ odd}, \quad (48.52)$$

$$= (\sum \lambda_i v_i - v) / (\sum v_i + v), \quad n \text{ even}. \quad (48.53)$$

48.8 The distributional problem was solved by R. L. Anderson (1942) to whom the foregoing is due. We will illustrate the method on the particular case $n = 6$ and then quote the general result.

We have, for $n = 6$,

$$r = \frac{\lambda_1 v_1 + \lambda_2 v_2 - v}{v_1 + v_2 + v}, \quad p_6 = v_1 + v_2 + v, \quad (48.54)$$

where

$$\lambda_1 = \cos \frac{2\pi}{6} = \frac{1}{2}, \quad \lambda_2 = -\frac{1}{2}. \quad (48.55)$$

The joint distribution of the v 's is given by

$$dF(v_1, v_2, v) = \frac{1}{4(2\pi)^{\frac{1}{2}}} v^{-\frac{1}{2}} e^{-\frac{1}{2}v} e^{-\frac{1}{2}v_1} e^{-\frac{1}{2}v_2} dv dv_1 dv_2 \quad (48.56)$$

$$= \frac{1}{4(2\pi)^{\frac{1}{2}}} v^{-\frac{1}{2}} e^{-\frac{1}{2}p_6} dv dv_1 dv_2. \quad (48.57)$$

We have to consider two cases,

$$\lambda_2 \leq r \leq \lambda_1$$

and

$$-1 \leq r \leq \lambda_2.$$

Note from (48.52-3) that $r \leq \lambda_1$, the larger of the roots in λ . Consider the first case. From (48.54) we have

$$v_2 = \frac{1}{\lambda_1 - \lambda_2} \{p_6(\lambda_1 - r) - v(1 + \lambda_1)\} \quad (48.58)$$

$$v_1 = \frac{1}{\lambda_1 - \lambda_2} \{p_6(r - \lambda_2) + v(1 + \lambda_2)\}. \quad (48.59)$$

The Jacobian of the transformation will be found to be

$$\frac{\partial(v_1, v_2)}{\partial(p_6, r)} = \frac{p_6}{\lambda_1 - \lambda_2}.$$

Hence, from (48.57) the distribution of p_6 , r and v is given by

$$dF = \frac{1}{4(2\pi)^{\frac{1}{2}}} \frac{v^{-\frac{1}{2}} p_6 e^{-\frac{1}{2}p_6}}{\lambda_1 - \lambda_2} dp_6 dr dv. \quad (48.60)$$

Integrating out for v , noting the limits determined by (48.58) and (48.59), we have

$$\begin{aligned} dF &= \frac{1}{4(2\pi)^{\frac{1}{2}}} \frac{p_6 e^{-\frac{1}{2}p_6}}{\lambda_1 - \lambda_2} \left[2\omega^{\frac{1}{2}} \right]_0^{p_6 \left(\frac{\lambda_1 - r}{1 + \lambda_1} \right)} dp_6 dr \\ &= \frac{1}{2(2\pi)^{\frac{1}{2}} (\lambda_1 - \lambda_2) (1 + \lambda_1)^{\frac{1}{2}}} p_6^{3/2} e^{-\frac{1}{2}p_6} (\lambda_1 - r)^{\frac{1}{2}} dp_6 dr. \end{aligned} \quad (48.61)$$

Finally, integrating for p_6 from 0 to ∞ , we obtain

$$dF(r) = \frac{3}{2} \frac{(\lambda_1 - r)^{\frac{1}{2}}}{(\lambda_1 - \lambda_2) (1 + \lambda_1)^{\frac{1}{2}}} dr, \quad \lambda_2 \leq r \leq \lambda_1. \quad (48.62)$$

For the other part of the range we find similarly

$$dF(r) = \left\{ \frac{3}{2} \frac{(\lambda_1 - r)^{\frac{1}{2}}}{(\lambda_1 - \lambda_2) (1 + \lambda_1)^{\frac{1}{2}}} + \frac{3}{2} \frac{(\lambda_2 - r)^{\frac{1}{2}}}{(\lambda_2 - \lambda_1) (1 + \lambda_2)^{\frac{1}{2}}} \right\} dr, \quad -1 \leq r \leq \lambda_2. \quad (48.63)$$

48.9 It is typical of these distributions that they split into separate analytical expressions for values of r between the critical roots λ . The frequency curves are continuous, but the derivatives not necessarily so. It is also typical that the distribution functions are easily written down. For example, corresponding to (48.62) and (48.63) we have

$$\text{Prob}(R > r) = \frac{(\lambda_1 - r)^{\frac{3}{2}}}{(\lambda_1 - \lambda_2) (1 + \lambda_1)^{\frac{1}{2}}}, \quad \lambda_2 \leq r \leq \lambda_1, \quad (48.64)$$

$$= \frac{(\lambda_1 - r)^{\frac{3}{2}}}{(\lambda_1 - \lambda_2) (1 + \lambda_1)^{\frac{1}{2}}} + \frac{(\lambda_2 - r)^{\frac{3}{2}}}{(\lambda_2 - \lambda_1) (1 + \lambda_2)^{\frac{1}{2}}}, \quad -1 \leq r \leq \lambda_2. \quad (48.65)$$

We quote the general form of the complement to the distribution function:

$$\text{Prob } (R > r) = \sum_{i=1}^m \frac{(\lambda_i - r)^{\frac{1}{2}(n-3)}}{\prod_{j=1, i \neq j}^m (\lambda_i - \lambda_j)}, \quad \lambda_{m+1} \leq r \leq \lambda_m, \quad n \text{ odd}, \quad (48.66)$$

$$= \sum_{i=1}^m \frac{(\lambda_i - r)^{\frac{1}{2}(n-3)}}{\prod_{j=1, i \neq j}^m (\lambda_i - \lambda_j) (1 + \lambda_i)^{\frac{1}{2}}}, \quad \lambda_{m+1} \leq r \leq \lambda_m, \quad n \text{ even}. \quad (48.67)$$

The frequency function falls into $\frac{1}{2}(n-1)$ or $\frac{1}{2}(n-2)$ pieces according to the parity of n .

48.10 We shall have to pass over rather cursorily a number of features of the distribution which are of mathematical rather than statistical interest.

- (a) For r_l with l not equal to unity the circulant has factors typified by $\lambda_k - \cos(2\pi lk/n)$. If l is prime to n the circulant is the same as before and the distribution remains unchanged. If not, it would seem that the analytical form is different.
- (b) There are other ways of obtaining a circulant without assuming a circular definition of the coefficient. For example, with $n = 2m$,

$$\frac{u_1 u_2 + \dots + u_{m-1} u_m + u_{m+1} u_{m+2} + \dots + u_{n-1} u_n}{\sum u^2}$$

will be found to have the characteristic property of paired roots in λ , and therefore follows Anderson's distribution. Other cases are given by Durbin and Watson (1950-1).

48.11 We proceed to derive the characteristic function of q and p . Taking temporarily s, t as the dummy variables, we have for the joint c.f. of q and p , the numerator and denominator of r ,

$$\begin{aligned} \phi(s, t) &\propto \int \exp \left[-\frac{1}{2} \left\{ \sum u^2 - 2it \sum (u - \bar{u})^2 - 2is(u_1 u_2 + \dots + u_n u_1 - n\bar{u}^2) \right\} \right] du \\ &= \Delta^{-\frac{1}{2}} \end{aligned} \quad (48.68)$$

where

$$\Delta = \begin{vmatrix} 1 - 2it \left(1 - \frac{1}{n}\right) + \frac{2is}{n}, & 1 + \frac{it}{n} & \dots & \frac{it}{n} - is + \frac{2}{n} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{it}{n} - is + \frac{2}{n} & 1 + \frac{it}{n} & 1 - 2it \left(1 - \frac{1}{n}\right) + \frac{2is}{n} & \cdot \end{vmatrix} \quad (48.69)$$

This is the same kind of circulant which we had at (48.45) and reduces as at (48.49) to

$$\Delta = \prod_{k=1}^{n-1} \left\{ 1 - 2i \left(t + s \cos \frac{2\pi k}{n} \right) \right\}. \quad (48.70)$$

Taking logarithms and identifying coefficients we get, for the cumulants of q and p ,

$$\kappa_{ij} = i!j! \sum_{k=1}^{n-1} \frac{2^{i+j-1}}{i+j} \binom{i+j}{i} \left(\cos \frac{2\pi k}{n} \right)^i \quad (48.71)$$

$$= 2^r r! \sum \left(\cos \frac{2\pi k}{n} \right)^i, \quad r = i+j-1. \quad (48.72)$$

Now we have

$$\sum \cos \frac{2\pi k}{n} = -1 = \sum \cos^3 \frac{2\pi k}{n} = \sum \cos^5 \frac{2\pi k}{n} = \text{etc.} \quad (48.73)$$

$$\sum \cos^2 \frac{2\pi k}{n} = \frac{1}{2}(n-2), \quad \sum \cos^4 \frac{2\pi k}{n} = \frac{1}{8}(3n-8). \quad (48.74)$$

Substituting, we find

$$\begin{aligned} \kappa_{10} &= -1; \quad \kappa_{01} = n-1; \quad \kappa_{20} = n-2; \quad \kappa_{11} = -2; \quad \kappa_{02} = 2(n-1); \\ \kappa_{30} &= -8; \quad \kappa_{21} = 4(n-2); \quad \kappa_{12} = -8; \quad \kappa_{03} = 8(n-1). \end{aligned} \quad (48.75)$$

It will be seen that κ_{20} and κ_{02} are of order n , whereas higher order cumulants are of no higher order. Thus, in standard measure, the higher order cumulants tend to zero, and the distribution accordingly tends to bivariate normality. Moreover κ_{11} in standard measure tends to zero, so that q and p are uncorrelated and hence, through normality, independent in the limit.

In fact, r and p are independent for normal variation and hence

$$\begin{aligned} E(r^m p^m) &= E(r^m) E(p^m) = E(q^m) \\ E(r^m) &= \frac{E(q^m)}{E(p^m)}. \end{aligned} \quad (48.76)$$

We can then evaluate the moments of r from (48.75), finding

$$\mu_1(r) = -\frac{1}{n-1}$$

$$\mu_2(r) = \frac{n(n-3)}{(n+1)(n-1)^2} \quad (48.77)$$

$$\mu_3(r) = \frac{2(2n-1)(n-6)}{(n-1)^3(n+1)(n+3)}. \quad (48.78)$$

The mean agrees with the exact non-circular result at (48.25), but the variance is only the same to order n^{-1} as the non-circular variance (48.40).

48.12 Dixon (1944) obtained an approximate form for Anderson's distribution by an ingenious smoothing of the characteristic function (48.68-70). Write temporarily

$$\alpha = 1 - 2it, \quad \beta = -2is, \quad \theta_k = 2\pi k/n. \quad (48.79)$$

We then have approximately

$$\begin{aligned} \phi(s, t) &\propto \prod_{k=1}^{n-1} (\alpha + \beta \cos \theta_k)^{-\frac{1}{2}} \\ &\doteq (\alpha + \beta)^{\frac{1}{2}} \exp \left[-\frac{1}{2} \log \sum_{k=1}^n (\alpha + \beta \cos \theta_k) \right] \\ &\doteq (\alpha + \beta)^{\frac{1}{2}} \exp \left[-\frac{n}{4\pi} \cdot \frac{2\pi}{n} \sum_{k=1}^n \log (\alpha + \beta \cos \theta_k) \right] \\ &\doteq (\alpha + \beta)^{\frac{1}{2}} \exp \left[-\frac{n}{4\pi} \int_0^{2\pi} \log (\alpha + \beta \cos \theta) d\theta \right] \\ &\doteq (\alpha + \beta)^{\frac{1}{2}} \exp \left[-\frac{1}{2} n \log \left\{ \frac{1}{2} (\alpha + (\alpha^2 - \beta^2)^{\frac{1}{2}}) \right\} \right] \\ &\doteq 2^{\frac{1}{2}n} (\alpha + \beta)^{\frac{1}{2}} \{ \alpha + (\alpha^2 - \beta^2)^{\frac{1}{2}} \}^{-\frac{1}{2}n}. \end{aligned} \quad (48.80)$$

By successive differentiation we can now obtain the moments of q and p . In fact, by differentiating for α and integrating for β we can find moments of the form $E(q^k/p^k)$. We find the unexpectedly simple forms

$$\mu'_1 = -\frac{1}{n-1} \quad (48.81)$$

$$\mu'_2 = \frac{1}{n+1} \quad (48.82)$$

$$\mu'_3 = -\frac{3}{(n-1)(n+3)} \quad (48.83)$$

$$\mu'_4 = \frac{3}{(n+1)(n+3)}. \quad (48.84)$$

It may be shown (Dixon, 1944) that these moments are, in fact, exact. The moments of r^2 may be seen to be, up to the $(\frac{1}{2}n)th$, those of the distribution

$$\frac{\Gamma(\frac{1}{2}n + \frac{1}{2})}{\Gamma(\frac{1}{2}n) \Gamma(\frac{1}{2})} r^{-1} (1-r^2)^{\frac{1}{2}(n-2)}. \quad (48.85)$$

Thus, from (16.62), the squared serial correlation has the same distribution, approximately, as the squared ordinary correlation coefficient in samples of $n+2$ from an uncorrelated normal population.

Dixon (1944) also treats the case when the series has known (zero) mean and the case of a coefficient of lag $l > 1$. The same result holds, with $n+1$ replacing n when the mean is known, up to $(n/2m)$ moments, where m is the largest common factor of l and n . Cf. Exercise 48.20.

48.13 Koopmans (1942) reached the same result by a different route. He expressed the c.f. of p and q as a contour integral and smoothed the values of λ , as above, by spreading them uniformly round a circle *before integrating*. This led him to the expression

$$h(\kappa) = \frac{(\frac{1}{2}n-1)2^{\frac{1}{2}n}}{\pi} \int_0^{\arccos r} (\cos \alpha - r)^{\frac{1}{2}n-\alpha} \sin \frac{1}{2}n\alpha \sin \alpha \, d\alpha. \quad (48.86)$$

Rubin (1945) evaluated the integral by showing, with the aid of a partial integration, that

$$\frac{\partial}{\partial r} h(r, n) = -nrh(r, n-2)$$

and proceeding by induction.

48.14 We had better pause at this point and review progress. Our large-sample results are explicit, but troublesome to apply; and in any case, large samples in time-series analysis are rare outside the domain of physics and meteorology. For small samples we have obtained exact results for a random series; but again, we should not, as a rule, wish to apply tests to a series until we were satisfied by simpler tests that it was not random. Moreover, our exact results depend on the assumption of normality and, for the most part, apply to circularly defined coefficients. Nevertheless they are very illuminating and provide a number of interesting further problems.

The Madow-Leipnik distribution

48.15 If the statistics t_1, t_2, \dots, t_k are sufficient for $\theta_1, \theta_2, \dots, \theta_k$ we have, in an obvious notation,

$$P(u_1, \dots, u_n | \theta) = Q(t | \theta)R(u)$$

and hence

$$P(t | \theta) = P(t | \theta_0) \frac{Q(t | \theta)}{Q(t | \theta_0)}. \quad (48.87)$$

Madow (1945) used this result to derive the approximate distribution of the serial correlations for the non-null case from those for the null case $\theta = \theta_0$.

Suppose that u_1, u_2, \dots, u_n have a joint normal distribution, with mean μ , of the following type:

$$\log L = \text{constant} - \frac{1}{2} \left[A \sum_{i=1}^n (u_i - \mu)^2 + 2B \sum_{i=1}^n (u_i - \mu)(u_{i+j} - \mu) \right]. \quad (48.88)$$

As before, we assume a circular definition of r , with q and p as its numerator and denominator. Then \bar{u} , p and q are sufficient for μ , A and B . We take as null the Anderson case with $A = 1$, $B = 0$ and hence find, from (48.87),

$$P(r, p | A, B) \propto P(r, p | 1, 0) \frac{e^{-\frac{1}{2}(Ap + 2Brp)}}{e^{-\frac{1}{2}p}}$$

and since r and p are independent, with p a multiple of a $\chi^2(n-1)$ variable this is

$$\propto p^{\frac{1}{2}(n-3)} e^{-\frac{1}{2}p(A+2Br)} f(r). \quad (48.89)$$

We integrate out from $p = 0$ to ∞ to get

$$P(r | A, B) \propto \frac{1}{(\frac{1}{2}A + Br)^{\frac{1}{2}(n-1)}} f(r). \quad (48.90)$$

Thus the non-null distribution differs from the null distribution in having the factor $(\frac{1}{2}A + Br)^{\frac{1}{2}(n-1)}$ adjoined to it. If it is known that the u 's have zero mean, (48.90) is modified so as to have $\frac{1}{2}n$ as the exponent in the denominator.

48.16 In particular, let u obey the Markoff relation

$$u_t = \rho u_{t-1} + \varepsilon_t.$$

We know (cf. (47.70)) that

$$\text{var } u = \frac{1}{1-\rho^2} \text{var } \varepsilon$$

and

$$\text{cov}(u_t, u_{t+1}) = \rho \text{var } u = \frac{\rho}{1-\rho^2} \text{var } \varepsilon.$$

Then the distribution of the ε 's has likelihood given by

$$\begin{aligned} \log L &= \text{constant} - \frac{1}{2\sigma^2} \sum_{i=1}^n \varepsilon_i^2 \\ &= \text{constant} - \frac{1}{2\sigma^2} \sum (u_i - \rho u_{i-1})^2 \end{aligned}$$

where $\sigma^2 = \text{var } \varepsilon$. Approximately this is equivalent to

$$\log L = \text{const.} - \frac{1}{2\sigma^2}(1+\rho^2) \sum u_i^2 + \frac{\rho}{\sigma^2} \sum u_i u_{i+1}. \quad (48.91)$$

Hence, in (48.88) we have

$$A = \frac{1+\rho^2}{\sigma^2}, \quad B = -\frac{\rho}{\sigma^2},$$

$$\frac{1}{2}A + Br \propto (1-2\rho r + \rho^2).$$

and thus

Thus, from (48.90) with the modified $\frac{1}{2}n$ in the denominator, and the remaining factor replaced by the approximate form (48.85), we have

$$P(r | \rho) = \frac{\Gamma\{\frac{1}{2}(n+1)\}}{\Gamma(\frac{1}{2}n + \frac{1}{2})\Gamma(\frac{1}{2})} \frac{(1-r^2)^{\frac{1}{2}(n-1)}}{(1-2\rho r + \rho^2)^{\frac{1}{2}n}}, \quad -1 \leq r \leq 1. \quad (48.93)$$

48.17 This remarkable form has been studied by Leipnik (1947), Quenouille (1948), Jenkins (1956) and Kendall (1957). Its moments are not nearly so easy to obtain by straightforward integration as might be expected.

For the moment of order k about the origin, writing C for a constant, we have

$$\begin{aligned} \mu'_k &= C \int_{-1}^1 \frac{(1-r^2)^{\frac{1}{2}(n-1)} r^k}{(1+\rho^2-2\rho r)^{\frac{1}{2}n}} dr \\ -\frac{1}{n} \frac{\partial \mu'_k}{\partial \rho} &= \int_{-1}^1 \frac{\rho r r^{k-1}}{(1+\rho^2-2\rho r)} dF - \int_{-1}^1 \frac{(1+\rho^2)-(1+\rho^2-2\rho r)}{2\rho(1+\rho^2-2\rho r)} dF \\ &= \frac{1}{2} \int \frac{(1+\rho^2)-(1+\rho^2-2\rho r)}{1+\rho^2-2\rho r} r^{k-1} dF - \frac{1+\rho^2}{2\rho} \int \frac{r^k dF}{1+\rho^2-2\rho r} + \frac{1}{2\rho} \mu'_k \\ &= \frac{1}{2\rho} \mu'_k - \frac{1}{2} \mu'_{k-1} + \frac{1+\rho^2}{2\rho} \int \frac{(\rho-r)r^{k-1}}{1+\rho^2-2\rho r} dF, \end{aligned}$$

whence we find

$$\left(\frac{1}{n} \frac{\partial}{\partial \rho} + \frac{1}{2\rho}\right) \mu'_k = \left(\frac{1+\rho^2}{2\rho n} \frac{\partial}{\partial \rho} + \frac{1}{2}\right) \mu'_{k-1}. \quad (48.94)$$

It will then be evident by induction that μ_k is a polynomial of order k in ρ . Moreover, even-order moments contain only even powers of ρ , and odd-order contain only odd-order powers of ρ .

Let

$$\mu'_k = \sum_{m=0}^k a_{km} \rho^m. \quad (48.95)$$

Differentiating (48.94) m times and putting $\rho = 0$ we find

$$a_{km} = \frac{1}{n+2m} \left\{ (m+1)a_{k-1, m+1} + (n+m-1)a_{k-1, m-1} \right\}. \quad (48.96)$$

We then find, since $\mu_0 = 1$, $a_{00} = 1$,

$$a_{11} = \frac{n}{n+2}$$

giving

$$\mu'_1 = \frac{n\rho}{n+2}. \quad (48.97)$$

Successive applications of (48.96) then yield

$$\mu'_2 = \frac{1}{n+2} + \frac{n(n+1)}{(n+2)(n+4)}\rho^2 \quad (48.98)$$

$$\mu'_3 = \frac{3n\rho}{(n+2)(n+4)} + \frac{n(n+1)}{(n+4)(n+6)}\rho^3 \quad (48.99)$$

$$\mu'_4 = \frac{3}{(n+2)(n+4)} + \frac{6n(n+1)}{(n+2)(n+4)(n+6)}\rho^2 + \frac{n(n+1)(n+3)}{(n+4)(n+6)(n+8)}\rho^4 \quad (48.100)$$

and so on. In particular, for the moments about the mean

$$\mu_2 = \frac{1}{n+2} - \frac{n(n-2)}{(n+2)^2(n+4)}\rho^2 \sim \frac{1-\rho^2}{n} \quad (48.101)$$

$$\mu_3 = \frac{-6n\rho}{(n+2)^2(n+4)} + \frac{2n(n-2)(3n-2)}{(n+2)^3(n+4)(n+6)}\rho^3 \sim \frac{-6\rho(1-\rho^2)}{n^2} \quad (48.102)$$

$$\mu_4 \sim \frac{3(1-\rho^2)^2}{n^2}. \quad (48.103)$$

We note in particular that, in standard measure, μ_3 tends to zero and μ_4 tends to 3, illustrating the tendency of the distribution to normality.

The distribution, it must be remembered, refers to the case when the mean of u is known (and can therefore be assumed to be zero).

White (1957) and Leipnik (1958) have derived expressions for the moments in terms of polynomials of the Gegenbauer type; Leipnik derives an expression for the characteristic function and shows that it tends to the normal form.

48.18 The approximate form of the variance given in (48.101)

$$\text{var } r = \frac{1-\rho^2}{n}$$

(which, we note in passing, is not the same form as for a product-moment coefficient) suggests the normalizing transformation

$$r = \sin z, \quad \rho = \sin \zeta.$$

This was tried by Jenkins (1954), who found for $x = z - \zeta$

$$\mu'_1(x) = -\frac{3}{2} \frac{\rho}{(1-\rho^2)^{\frac{1}{2}}} \cdot \frac{1}{n} + O(n^{-2}) \quad (48.104)$$

$$\mu'_2(x) = \frac{1}{n} - \frac{2-5\rho^2}{2(1-\rho^2)} \cdot \frac{1}{n^2} + O(n^{-3}) \quad (48.105)$$

$$\gamma_1 = \frac{-3\rho}{(1-\rho^2)^{\frac{1}{2}}} n^{-\frac{1}{2}} + O(n^{-1}) \quad (48.106)$$

$$\gamma_2 = \frac{2(7\rho^2-1)}{1-\rho^2} n^{-1} + O(n^{-2}). \quad (48.107)$$

For moderate values of ρ this may be adequate, but evidently breaks down near $\rho = 1$.

Daniels' approximations

48.19 A major advance in the derivation of sampling distributions of serial correlations was made by Daniels (1956), who had introduced saddlepoint methods into statistics systematically for the first time. For a detailed account of those methods we refer to Jeffreys and Jeffreys' book on *The Methods of Mathematical Physics* (Oxford, 1956). Briefly, they amount to this: in the complex plane the integral of an analytic function around a contour which contains no singularities is zero; thus one path of integration can be deformed into another, provided that no singularity is crossed. The method looks for a path which runs through a saddlepoint of the surface, the presumption being that there the function falls away most steeply from its maximum, and hence that, in the neighbourhood of the saddlepoint, the values of the function being integrated are most highly concentrated. The path then gives us the steepest descent from a maximum value, and an expansion round the point will give us a good approximation to the integral required.

In applications the method simplifies where means or ratios of mean quantities are concerned.

48.20 As before, let $r = q/p$. Let $M(\theta_1, \theta_2)$ be the moment-generating function of p and q , i.e.

$$M(\theta_1, \theta_2) = \int e^{\theta_1 p + \theta_2 q} dF. \quad (48.108)$$

In terms of the Fourier inversion we have

$$f(p, q) = \frac{1}{(2\pi i)^2} \iint M(\theta_1, \theta_2) e^{-\theta_1 p - \theta_2 q} d\theta_1 d\theta_2, \quad (48.109)$$

the integration being along the *imaginary* axes of θ_1, θ_2 . In particular,

$$f(p, rp) = \frac{1}{(2\pi i)^2} \iint M(\theta_3 - r\theta_2, \theta_2) e^{-p\theta_3} d\theta_3 d\theta_2,$$

where the integration of $\theta_3 = \theta_1 + r\theta_2$ is taken over the imaginary axis in the θ_3 plane, or any deformation of it which is permissible. Inverting with respect to θ_3 , we have

$$\int_0^\infty f(p, rp) e^{\theta_3 p} dp = \frac{1}{2\pi i} \int M(\theta_3 + r\theta_2, \theta_2) d\theta_2 \quad (48.110)$$

so that, when differentiation is permissible,

$$\int_0^\infty p f(p, rp) e^{\theta_3 p} dp = \frac{1}{2\pi i} \int \frac{\partial M(\theta_3 - r\theta_2, \theta_2)}{\partial \theta_3} d\theta_2, \quad (48.111)$$

and, since the expression on the left with $\theta_3 = 0$ is the frequency distribution of r , we have

$$h(r) = \frac{1}{2\pi i} \int \frac{\partial M(\theta_3 - r\theta_2, \theta_2)}{\partial \theta_3} \Big|_{\theta_3=0} d\theta_2. \quad (48.112)$$

This is equivalent to Geary's result of (11.78), Vol. 1, otherwise proved in Exercise 11.24. Frequently we wish to transform q , and hence θ_2 , to some other form. If the transformation is such that $\theta_2 = \theta_2(\theta_4, \theta_3)$, then (48.112) becomes

$$h(r) = \frac{1}{2\pi i} \int \frac{\partial}{\partial \theta_3} \left[\{M(\theta_3 - r\theta_2, \theta_2)\} \frac{\partial \theta_2}{\partial \theta_4} \right]_{\theta_3=0} d\theta_4. \quad (48.113)$$

48.21 Consider first of all a coefficient circularly defined, with known (zero) mean and unit variance, from a circular Markoff process. The joint distribution of the u 's is then (cf. (48.91))

$$dF = \frac{1-\rho^n}{(2\pi)^{\frac{1}{2}n}} \exp \left[-\frac{1}{2} \left\{ (1+\rho^2) \sum_{i=1}^n u_i^2 - 2\rho \sum_{i=1}^n u_i u_{i+1} \right\} \right] du_1 \dots du_n, \quad (48.114)$$

with, of course, $u_1 = u_n$.

The m.g.f. (compare (48.69)) is

$$M(\theta_1, \theta_2) = (1-\rho^n) \Delta^{-\frac{1}{2}} \quad (48.115)$$

where

$$\Delta = \begin{vmatrix} 1+\rho^2-2\theta_1 & -(\rho+\theta_2) & 0 & \dots & -(\rho+\theta_2) \\ -(\rho+\theta_2) & 1+\rho^2-2\theta_1 & -(\rho+\theta_2) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -(\rho+\theta_2) & 0 & 0 & \dots & 1+\rho^2-2\theta_1 \end{vmatrix} \quad (48.116)$$

This is a circulant which reduces to

$$\Delta = \{1+\rho^2-2\theta_1-2(\rho+\theta_2)\} \prod_{k=1}^{n-1} \left\{ 1+\rho^2-2\theta_1-2(\rho+\theta_2) \cos \frac{2\pi k}{n} \right\}.$$

Put

$$\frac{1+\rho^2-2\theta_1}{\rho+\theta_2} = z + \frac{1}{z}. \quad (48.117)$$

Then Δ reduces to

$$\begin{aligned} \Delta &= \frac{(\rho+\theta_2)^n}{z^n} \prod \left(z^2 - 2z \cos \frac{2\pi k}{n} + 1 \right) \\ &= \frac{(\rho+\theta_2)^n (1-z^n)^2}{z^n}. \end{aligned} \quad (48.118)$$

Hence

$$M(\theta_3 - r\theta_2, \theta_2) = \frac{1-\rho^n}{1-z^n} \frac{(1-2rz+z^2)^{\frac{1}{2}n}}{(1-2\rho r+\rho^2-2\theta_3)^{\frac{1}{2}n}}, \quad (48.119)$$

where

$$\theta_2 = -\rho + \frac{z(1-2\rho r+\rho^2-2\theta_3)}{1-2rz+z^2}. \quad (48.120)$$

Then

$$\left(\frac{\partial \theta_2}{\partial z} \right)_{\theta_3} M = \frac{(1-\rho^n)(1-z^2)(1-2rz+z^2)^{\frac{1}{2}n-2}}{(1-z^n)(1-2\rho r+\rho^2-2\theta_3)^{\frac{1}{2}n-1}} \quad (48.121)$$

and (48.113) becomes

$$h(r) = \frac{(n-2)(1-\rho^n)}{2\pi i (1-2\rho r+\rho^2)^{\frac{1}{2}n}} \int \frac{(1-z^2)(1-2rz+z^2)^{\frac{1}{2}n-2}}{1-z^n} dz. \quad (48.122)$$

There remains to determine the path of integration. Consider the pair of transformations which together compose (48.117):

$$z + \frac{1}{z} = \zeta, \quad \zeta - r = \frac{1-2\rho r+\rho^2}{2(\rho+\theta_2)}. \quad (48.123)$$

The region $|z| \leq 1$ will be seen to be mapped on the whole θ_2 -plane cut along the real axis exterior to the interval

$$-\frac{(1+\rho)^2}{2(1+r)}, \quad \frac{(1-\rho)^2}{2(1-r)}.$$

Any path in the θ_2 -plane running from $\tau - i\infty$ through the gap in the real axis to $\tau' + i\infty$ corresponds to a path in $|z| \leq 1$ running from $e^{-i\phi}$ to $e^{i\phi}$, where $r = \cos \phi$. The path of integration for (48.122) is therefore of this form.

Consider the factor $(1 - 2rz + z^2)^{\frac{1}{2}n-2} = \{1 - r^2 - (z-r)^2\}^{\frac{1}{2}n-2}$ in the integrand of (48.122). It has a saddlepoint (maximum) where $z = r$, a real point. Further, the path of steepest descent away from this point is perpendicular to the real axis (a result we quote without proof). Thus the path of integration required is the straight line joining $e^{-i\phi}$, $e^{i\phi}$.

So far the results are exact. Let us now neglect the factor ρ^n in (48.122) and $1 - z^n$ in the denominator of the integrand in (48.122). We then have

$$h(r) \sim \frac{n-2}{2\pi i(1-2\rho r + \rho^2)^{\frac{1}{2}n}} \int (1-z^2)(1-2rz+z^2)^{\frac{1}{2}n-2} dz. \quad (48.124)$$

$z = r + iw(1-r^2)^{\frac{1}{2}}$

Put

Then $-1 \leq w \leq 1$ and we have

$$\begin{aligned} h(r) &= \frac{(n-2)(1-r^2)^{\frac{1}{2}(n-1)}}{2\pi(1-2\rho r + \rho^2)^{\frac{1}{2}n}} \int_{-1}^1 (1+w^2)(1-w^2)^{\frac{1}{2}n-2} dw \\ &= \frac{\Gamma(\frac{1}{2}n+1)(1-r^2)^{\frac{1}{2}(n-1)}}{\pi^{\frac{1}{2}}\Gamma(\frac{1}{2}n+\frac{1}{2})(1-2\rho r + \rho^2)^{\frac{1}{2}n}}, \end{aligned} \quad (48.125)$$

which is the Madow-Leipnik distribution (48.93).

48.22 If we are not content to neglect the factor $1 - z^n$ in (48.122) the integral cannot be evaluated in closed form. If, however, we expand it in powers of z^n we obtain on integration the series

$$\begin{aligned} h(r) &= \frac{\Gamma(\frac{1}{2}n+1)(1-\rho^n)}{\pi^{\frac{1}{2}}(1-2\rho r + \rho^2)^{\frac{1}{2}n}} \left\{ \frac{(1-r^2)^{\frac{1}{2}(n-1)}}{\Gamma(\frac{1}{2}n+\frac{1}{2})} - \frac{3}{2^n \Gamma(\frac{3}{2}n+\frac{1}{2})} \frac{d^n}{dr^n} (1-r^2)^{\frac{1}{2}(3n-1)} \right. \\ &\quad \left. + \frac{5}{2^{2n} \Gamma(\frac{5}{2}n+\frac{1}{2})} \frac{d^{2n}}{dr^{2n}} (1-r^2)^{\frac{1}{2}(5n-1)} - \dots \right\} \end{aligned} \quad (48.126)$$

Daniels (1956), to whom this result is due, has obtained an upper bound to the error involved in approximating to (48.122) by its first term. The error is, in fact, small near $\rho = 0$ but not near $\rho = 0.5$ for n of the order of 20.

48.23 By the use of this method Daniels obtained a number of further results which we quote without the detailed derivation.

In the Markoff case, for a circular process with unknown mean and a circular serial correlation coefficient,

$$h(r) = \frac{(n-3)(1-\rho^n)}{2\pi i(1-\rho)(1-2\rho r + \rho^2)^{\frac{1}{2}(n-1)}} \int \frac{(1-z)(1-z^2)(1-2rz+z^2)^{\frac{1}{2}(n-5)}}{1-z^n}. \quad (48.127)$$

Again ignoring the factor z^n we have

$$h(r) = \frac{n\Gamma(\frac{1}{2}n-\frac{1}{2})(1-r^2)^{\frac{1}{2}n-1}}{2\pi^{\frac{1}{2}}\Gamma(\frac{1}{2}n)(1-\rho)(1-2\rho r + \rho^2)^{\frac{1}{2}(n-1)}} \left\{ 1 - r - \frac{1}{n}(1+r) \right\}. \quad (48.128)$$

It is to be noted that we can derive the moments of this distribution from those of the Madow-Leipnik distribution (48.93).

An alternative form may be derived by the following considerations: the final term in (48.128) is relatively of order n^{-1} . If we replace r by ρ in it, the result remains true to order n^{-1} . We may therefore remove it, but the constants in (48.128) then need revaluation to ensure that the integral of $h(r)$ over the variate range is unity. We arrive at

$$h(r) = \frac{\Gamma(\frac{1}{2}n + \frac{3}{2})}{2\pi^{\frac{1}{2}} \Gamma(\frac{1}{2}n)} \cdot \frac{(1-r)(1-r^2)^{\frac{1}{2}n-1}}{(1-2\rho r + \rho^2)^{\frac{1}{2}(n-1)}} \{1 + O(n^{-3/2})\}. \quad (48.129)$$

48.24 For the Markoff case and a non-circular process with known mean consider the non-circular statistic, defined by

$$r = \frac{u_1 u_2 + \dots + u_{n-1} u_n}{\frac{1}{2}u_1^2 + u_2^2 + \dots + u_{n-1}^2 + \frac{1}{2}u_n^2}. \quad (48.130)$$

Daniels finds

$$h(r) = \frac{n}{2\pi^{\frac{1}{2}}} \frac{\Gamma(\frac{1}{2}n-1)}{\Gamma(\frac{1}{2}n-2)} \frac{(1-\rho^2)^{\frac{1}{2}}(1-r^2)^{\frac{1}{2}n-1}}{(1-\rho r)(1-2\rho r + \rho^2)^{\frac{1}{2}n-1}} \{1 + O(n^{-1})\} \quad (48.131)$$

or an equivalent approximation

$$h(r) = \frac{\Gamma(\frac{1}{2}N+1)}{\pi^{\frac{1}{2}} \Gamma(\frac{1}{2}N+2)} \frac{(1-r^2)^{\frac{1}{2}(N-1)}}{(1-2\rho r + \rho^2)^{\frac{1}{2}N}} \{1 + O(n^{-3/2})\} \quad (48.132)$$

where

$$N = n-1 + \frac{\rho^2}{1-\rho^2}. \quad (48.133)$$

48.25 For the non-circular Markoff process with unknown mean, using r defined by (48.130) and N by (48.133) we have

$$h(r) = \frac{\Gamma(\frac{1}{2}N + \frac{3}{2})}{2\pi^{\frac{1}{2}} \Gamma(\frac{1}{2}N)} \frac{(1-r)(1-r^2)^{\frac{1}{2}N-1}}{\{N(1-\rho) - (1+\rho)\} (1-2\rho r + \rho^2)^{\frac{1}{2}(N-1)}} \{1 + O(n^{-3/2})\}. \quad (48.134)$$

48.26 It is possible to take these results a good deal further. Daniels (1956) deals with the general autoregressive process circularly defined.

In cases of higher order than the Markoff process it is also of some interest to consider the joint distribution of two or more serial correlations and of the partial correlations. Quenouille (1949b) was the first to do so. For some later work see Jenkins (1954, 1956), Watson (1956) and Daniels (1956). To save space we shall not enter into a detailed discussion of the results. Except in the Markoff case it appears that only circularly defined statistics and processes are reasonably tractable. Daniels' method can be extended to the non-circular case, but apparently nobody has yet had the stamina to embark on the labour involved.

48.27 A final word may be added concerning distributions when the residual term ε in an autoregressive scheme is not normal. The matter has not received much attention from statisticians. Quenouille (1948) did some sampling experiments, in particular with rectangularly distributed ε , and came to the conclusion that approximate normal theory provides satisfactory tests of serial correlations, at least for moderate or large samples.

THE ADVANCED THEORY OF STATISTICS

EXERCISES

48.1 Verify equation (48.12).

48.2 In a Yule scheme given by

$$u_t - 1.1u_{t-1} + 0.5u_{t-2} = \varepsilon_t$$

show that approximately

$$\text{var } r_1 = 2.44/n.$$

(Bartlett, 1946)

48.3 For a series in which $\rho_s = 0$, $s \geq 4$, show that

$$E(r_j) = -\frac{1}{n-j} (1 + 2\rho_1 + 2\rho_2 + 2\rho_3), \quad j > 3.$$

48.4 A sample correlation coefficient is defined circularly as

$$r_1 = \frac{\sum_{i=1}^n u_i u_{i+1} - n\bar{u}^2}{\sum_{i=1}^n u_i^2 - n\bar{u}^2}.$$

Show that in a Markoff scheme

$$E(r_1) = \rho - \frac{1 + 4\rho}{n}.$$

(Kendall, 1954)

48.5 For the circular definition of the previous exercise show that, in samples from a normal random series,

$$\text{var } (r_1) = \frac{n(n-3)}{(n+1)(n-1)^2}.$$

(Moran, 1947)

48.6 In continuation of 48.6 show that

$$\mu_3(r_1) = \frac{4(n-4)(n-3)}{(n-1)^4 (n+3)}$$

and that, for the circular definition,

$$\mu_3 = \frac{2(2n-1)(n-6)}{(n-1)^3 (n+1)(n+3)}.$$

(Moran, 1947)

48.7 Evaluate (48.9) for the Markoff scheme and hence reconcile it with (48.101).

48.8 For a large sample from a Markoff process, in the manner of 48.1, show that, in respect of r_j ,

$$\begin{aligned} \text{var } v &= \frac{(\kappa_4 + 2)(1 + \rho^2)}{n(1 - \rho^2)} \\ \text{var } c &= \frac{1}{n} \left[\frac{(1 + \rho^2)(1 - \rho^{2j})}{1 - \rho^2} + \rho^{2j} \left\{ 2j + \frac{(\kappa_4 + 2)(1 + \rho^2)}{1 - \rho^2} \right\} \right] \\ \text{cov } (c, v) &= \frac{\rho^j}{n} \left[2j + \frac{(\kappa_4 + 2)(1 + \rho^2)}{1 - \rho^2} \right] \end{aligned}$$

and hence that, independently of κ_4 ,

$$\text{var } r_j = \frac{1}{n} \left[\frac{(1+\rho^2)(1-\rho^{2j})}{1-\rho^2} - 2j\rho^{2j} \right]$$

(Bartlett, 1946)

48.9 Analogously to (48.9), show that for large samples,

$$\text{cov}(r_j, r_{j+k}) = \frac{1}{n} \sum_{i=-\infty}^{\infty} \{ \rho_i \rho_{i+j} + \rho_i \rho_{i+2j+k} + 4\rho_j \rho_{j+k} \rho_i^2 - 2\rho_j \rho_i \rho_{i+j+k} - 2\rho_{j+k} \rho_i \rho_{i+k} \}.$$

(Bartlett, 1946)

48.10 Show (cf. Example 48.4) that the large-sample variance of r_j at (48.9) is given by

$$n \text{ var } r_j = [(1+2\rho_j^2) \text{co. } z^0 + \text{co. } z^{2j} - 4\rho_j \text{co. } z^j] \text{ in } G^2(z)$$

where $G(z)$ is the autocorrelation generating function and "co" means "the coefficient of". Verify on the result of Exercise 48.8.

48.11 Continuing Exercise 48.4, show that

$$E(r_j) = \rho^j - \frac{1}{n} \left\{ \frac{1+\rho}{1-\rho} (1-\rho^j) + 3j\rho^j - j\rho^{n-j} \right\}.$$

(Kendall, 1954)

48.12 Defining the first serial correlation (known zero mean) as

$$r_1 = \frac{n}{n-1} \frac{\sum_{i=1}^{n-1} u_i u_{i+1}}{\sum_{i=1}^n u_i^2}$$

show that, for a random series, r_1 and the denominator are independent and hence derive

$$\mu_1(r_1) = 0 = \mu_3(r_1)$$

$$\mu_2(r_1) = \frac{n}{(n-1)(n+2)}$$

$$\mu_4(r_1) = \frac{3n^3(n^2+4n-9)}{(n-1)^4(n+2)(n+4)(n+6)}.$$

(Moran, 1948)

48.13 Starting from the Madow-Leipnik distribution of (48.93), make the transformation

$$r = \tanh z, \quad \rho = \tanh \zeta, \quad z - \zeta = x,$$

and by expansion show that the distribution of x is given by

$$h(x) = \frac{1}{\sigma\sqrt{2n}} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right) X$$

$$\sigma^2 = \cosh^2 \zeta / n = 1 / \{n(1-\rho^2)\}$$

where
and

$$X = 1 - \rho x + \left\{ \frac{1}{4n} - \frac{x^2(1-\rho^2)}{2} + \frac{nx^4(1-\rho^2)(1-3\rho^2)}{12} + \frac{\rho^2 x^2}{2} \right\} \\ - \rho x \left\{ \frac{1}{4n} - \frac{5x^2(1-\rho^2)}{6} + \frac{nx^4(1-\rho^2)(1-3\rho^2)}{12} + \frac{\rho^2 x^2}{6} \right\} + O(n^{-2}).$$

Obtain the moments and show that z is distributed approximately normally about mean

$$\zeta - \frac{\rho}{n(1-\rho^2)} + \frac{\rho(1+\rho^2)}{n^2(1-\rho^2)^2}$$

with variance

$$\frac{1}{n(1-\rho^2)} - \frac{2\rho^2}{n^2(1-\rho^2)^2}.$$

(Quenouille, 1948)

48.14 Taking the Yule scheme of Exercise 48.2, show that a generating function for the autocorrelations is given by

$$\begin{aligned} \sigma^2 \sum_{-\infty}^{\infty} \rho_i z^i &= \frac{1}{(1-1.1z+0.5z^2)(1-1.1z^{-1}+0.5z^{-2})} \\ &= 1 + \frac{0.7333-0.5z}{1-1.1z+0.5z^2} z + \frac{0.7333-0.5z^{-1}}{1-1.1z^{-1}+0.5z^{-2}} z^{-1} \end{aligned}$$

where $\sigma^2 = \text{var } u$. Squaring and expanding, show that

$$\sum_{-\infty}^{\infty} \rho_i^2 = 2.44,$$

and hence confirm that approximately, for large samples,
 $\text{var } r_j = 2.44/n$.

(Quenouille, 1947a)

48.15 For the general linear autoregressive scheme $\sum_{j=1}^k \alpha_j u_{t-j} = \varepsilon_t$ show that

$$\lim_{m \rightarrow \infty} \frac{\text{var } \Delta^m u_t}{\binom{2m}{m}} = \text{var } \varepsilon \left\{ \sum_{j=1}^k (-1)^j \alpha_j \right\}^{-2}.$$

(Murteira, 1951)

48.16 For the scheme of the previous exercise, in which ε_t is replaced by $\sum_{l=1}^p \beta_l \varepsilon_{t-l}$, show that the same limit takes the value

$$\text{var } \varepsilon \left\{ \frac{\sum_{i=1}^p (-1)^i \beta_i}{\sum_{j=1}^p (-1)^j \alpha_j} \right\}^2.$$

(Murteira, 1951)

48.17 In the Madow-Leipnik distribution (48.93), put

$$r = \sin y, \quad \rho = \sin \lambda, \quad y - \lambda = x,$$

and show that

$$\begin{aligned} h(x) = & \sqrt{\frac{n}{2\pi}} \exp(-\frac{1}{2}nx^2) \left\{ 1 - \frac{np}{2(1-\rho^2)^{\frac{1}{2}}} x^3 + \frac{1}{4n} - \frac{1}{24} \frac{(2+13\rho^2)}{1-\rho^2} nx^4 + \frac{1}{8} \frac{\rho^2}{1-\rho^2} n^2 x^6 - \frac{1}{8} \frac{\rho(1+5\rho^2)}{(1-\rho^2)^{3/2}} nx^5 \right. \\ & \left. - \frac{1}{8} \frac{\rho}{(1-\rho^2)^{\frac{1}{2}}} x^3 + \frac{1}{48} \frac{\rho(2+13\rho^2)}{(1-\rho^2)^{3/2}} n^2 x^7 - \frac{1}{48} \frac{\rho^3}{(1-\rho^2)^{3/2}} n^3 x^9 + O(n^{-2}) \right\} \end{aligned}$$

Hence derive equations (48.104) and (48.105).

(Jenkins, 1954)

48.18 In samples from a random series with zero mean and unit variance, show that the characteristic function of

$$p = \frac{1}{2} \sum u_i^2, \quad q = \frac{1}{2} \sum u_i u_{i+1}, \quad r = \frac{1}{2} \sum u_i u_{i+2},$$

circularly defined, is the circulant

$$\prod_{k=1}^n \left(1 - \theta_p - \theta_q \cos \frac{2\pi k}{n} - \theta_r \cos \frac{4\pi k}{n} \right)^{-\frac{1}{2}}.$$

Deduce that if

$$\mu_{st} = E \left(\frac{q^s r^t}{p^{s+t}} \right)$$

$$\mu_{10} = \mu_{11} = 0, \quad \mu_{20} = \mu_{02} = \frac{1}{n+2}$$

$$\mu_{11} = \mu_{12} = \mu_{21} = 0$$

$$\mu_{21} = \frac{2}{(n+2)(n+4)}$$

$$\mu_{22} = \frac{n+12}{(n+2)(n+4)(n+6)}.$$

(Jenkins, 1954)

48.19 Following the previous exercise, if statistics are defined with mean \bar{u} , e.g.

$$p = \frac{1}{2} \sum (u_i - \bar{u})^2,$$

show that the characteristic function of p, q, r is now

$$\prod_{k=1}^{n-1} \left(1 - \theta_p - \theta_q \cos \frac{2\pi k}{n} - \theta_r \cos \frac{4\pi k}{n} \right)^{-\frac{1}{2}}$$

and hence that

$$\mu_{11} = -\frac{1}{n^2-1}.$$

(Jenkins, 1954, who gives values for higher moments.)

48.20 For the statistic r_1 defined as in (48.42) but with known mean, and therefore with the omission of \bar{u} , show that the c.f., corresponding to (48.80), is the same with the omission of the factor in $(\alpha + \beta)^{\frac{1}{2}}$. Hence show that odd-order moments vanish and approximately

$$\mu_{2k} = \frac{1.3.5 \dots (2k-1)}{(n+2)(n+4) \dots (n+2k)}.$$

Hence verify (48.85) with $(n+1)$ replacing n .

(Dixon, 1944)

CHAPTER 49

SPECTRUM THEORY

Harmonic analysis

49.1 In Chapter 47 we have encountered the spectrum and associated functions as transforms of the autocorrelation function, but pointed out that they arose naturally from a different viewpoint, namely as measures of the closeness of the correlation between a time-series and certain harmonic terms. We proceed to develop this approach more fully.

Fourier was led by his studies of heat-flow to consider the expansion of functions in series of harmonic terms of the type

$$f(x) = \sum_{r=1}^{\infty} a_r \sin rx + \frac{1}{2}b_0 + \sum_{r=1}^{\infty} b_r \cos rx. \quad (49.1)$$

Notwithstanding the cyclical character of the individual terms, a very wide class of non-cyclical functions can be represented in this way over a limited range. It is, for example, sufficient that, in the range $-\pi$ to $+\pi$, $f(x)$ be single-valued, continuous except for a finite number of discontinuities, and have only a finite number of maxima or minima, for such an expansion to be valid. The series on the right in (49.1) is called a *Fourier series*. It has the attractive property that successive terms are orthogonal. For

$$\begin{aligned} \int_{-\pi}^{\pi} \cos rx \sin sx \, dx &= \int_{-\pi}^{\pi} \sin rx \sin sx \, dx \\ &= 0, \quad r \neq s, \\ &= \pi, \quad r = s. \end{aligned} \quad (49.2)$$

Hence, on multiplying (49.1) by $\sin rx$ and by $\cos rx$ and integrating, we find

$$a_r = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin rx \, dx, \quad (49.3)$$

$$b_r = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos rx \, dx. \quad (49.4)$$

The series may also be written in the form

$$f(x) = \sum_{r=0}^{\infty} c_r \sin (rx + \phi_r) \quad (49.5)$$

where ϕ_r is a phase angle.

Since all the terms in (49.1), apart from the constant, are of period 2π , the expression for $f(x)$ has that period. If $f(x)$ is defined over an interval $-L$ to L we may expand it in terms of $\sin(\pi rx/L)$ and $\cos(\pi rx/L)$. This, of course, is merely a matter of re-scaling the interval from one of length 2π to one of length $2L$.

49.2 Angles, measured as usual in radians, have zero dimensions. Thus the quantity αt in $\sin \alpha t$ has zero dimension and α is accordingly in radians per time-unit

It is sometimes called *angular frequency*. Where no ambiguity is involved we shall simply call it the frequency. However, $\sin \alpha t$ repeats itself with period $2\pi/\alpha$ and therefore the number of cycles per time-unit is $\alpha/2\pi$, which may also be regarded as the frequency. The period $2\pi/\alpha$ is of dimension t and is also called the "wavelength," although, in our context, "length" is a period of time.

49.3 It appears, then, that a function may be expanded in a series of sines and cosines, the successive terms in (49.1) having periods 2π , $2\pi/2$, $2\pi/3$, etc., and the corresponding angular frequencies being 1, 2, 3, with cycle frequencies $1/2\pi$, $2/2\pi$, $3/3\pi$. More generally, when $f(x)$ is defined over the interval $2L$, the angular frequencies are typified by π/L . Thus there is one fundamental frequency π/L and the others are integral multiples of it. Such a representation would be rather artificial if we knew that $f(x)$ was the sum of harmonic components with incommensurable frequencies. We are thus led to consider the more general harmonic series

$$f(x) = \sum_{j=0}^{\infty} a_j \sin(\alpha_j x) + \sum_{j=0}^{\infty} b_j \cos(\alpha_j x), \quad (49.6)$$

where the α 's can have any real values. There is now no simple way of evaluating a_j and b_j such as is given by (49.3) and (49.4). The problem of estimating them was considered in the nineteenth century by physicists and meteorologists, and although a great deal of knowledge has now been accumulated, the methods in essence are the same as those used by earlier authors. However, there has been a change of outlook. Former authors were looking for concealed harmonics. The more modern approach is to regard the spectrum as a characteristic of the time-series whether it is truly a sum of harmonics or not.

Nyquist frequency and aliases

49.4 For series observed at equal unit intervals of time there are two important features of harmonic analysis to observe. It is clearly possible for periodicities of less than one unit to escape notice—for example, if we observe a series every January 1st, seasonal movements will not be revealed. We need at least two observations in the year to detect periodicities of one year. Generally, for a time-interval t_0 between observations we cannot measure periods smaller than $2t_0$, or angular frequencies higher than π/t_0 . This limiting value is known as the *Nyquist frequency*.

In the spectral density function defined in 47.10 as

$$w(\alpha) = \sum_{-\infty}^{\infty} \rho_j e^{i\alpha j}, \quad (49.7)$$

our time-interval was unity and the range of α is from 0 to π . The ordinate at π represents the value of the spectral density at the Nyquist frequency.

49.5 The second effect to remark is also related to the interval of observation. Suppose that the interval is unity, and consider the term $\sin(2\pi t/3)$ for $t = 1, 2, 3$, etc. Its values are $\sqrt{3}/2$, $-\sqrt{3}/2$, 0, $\sqrt{3}/2$, etc. But these are also the values which would be observed for $\sin(8\pi t/3)$ or $\sin(14\pi t/3)$, etc. The width of the interval of observation does not permit of a distinction between angular frequency $2\pi/3$ or any of the

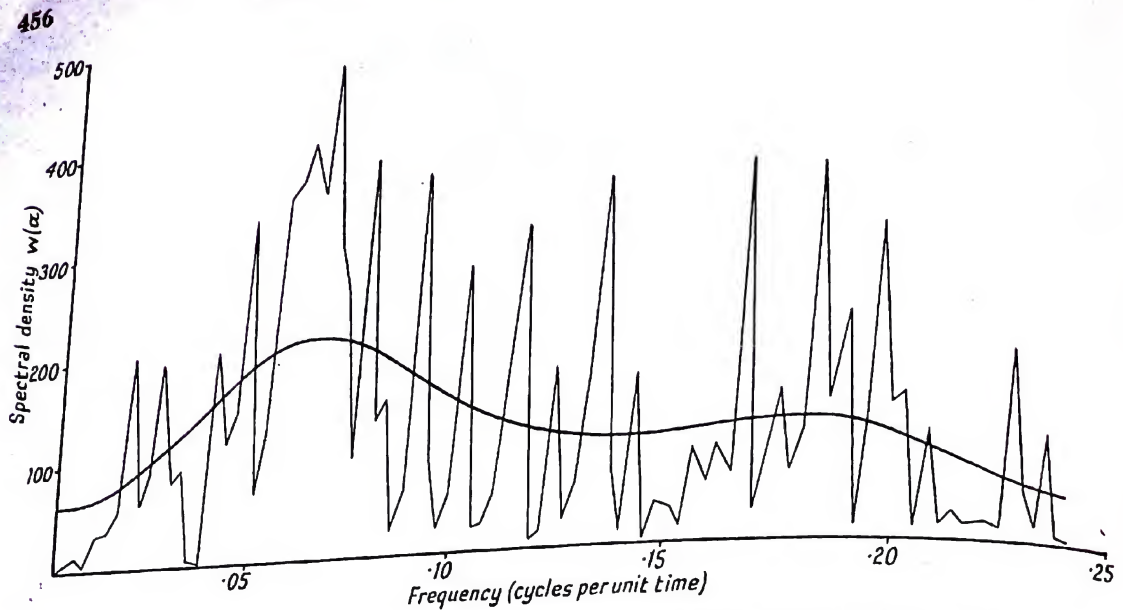


Fig. 49.1—Power spectrum of the Beveridge wheat-price index series (Table 47.1)
For clarity, only frequencies up to 0.24 are included, the remaining part of the spectrum being negligibly small. The curve is a smoothed spectrum using a Parzen kernel (Exercise 49.7). See also Fig. 49.4.

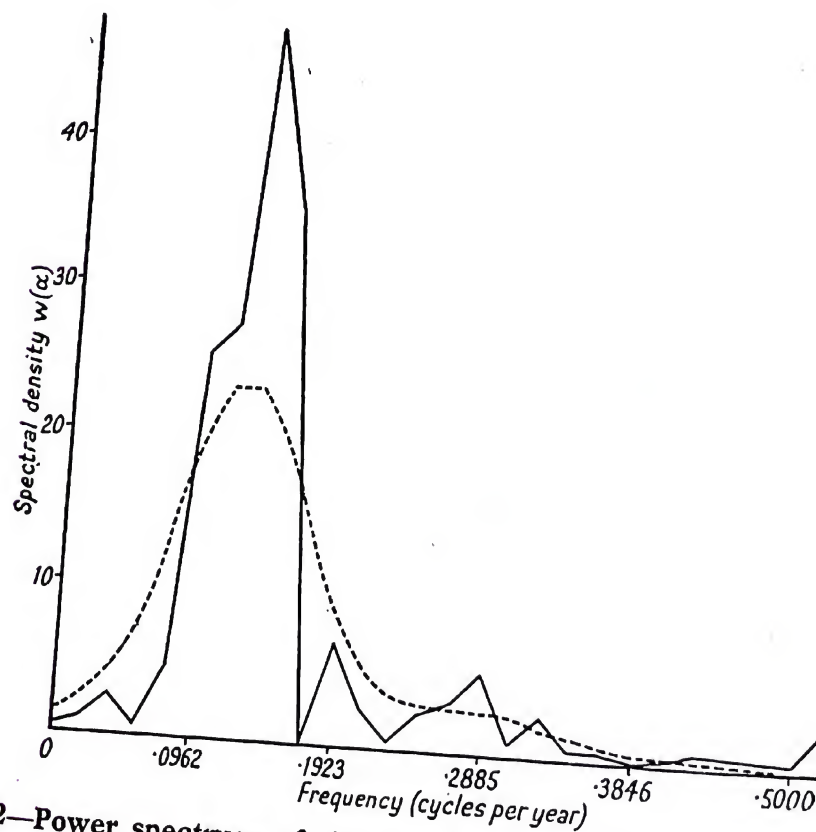


Fig. 49.2—Power spectrum of the data of Table 47.2 (marriage rates)
Maximum frequency at 0.1346 corresponding to a period of 7.4 years. The dotted line is a smoothed spectrum using a Parzen window (Exercise 49.7) and the first fifteen serial correlations.

angular frequencies $2\pi/3 + 2\pi j$, $j = 1, 2, 3$, etc. These higher frequencies are known as *aliases*. So far as observation goes they are all equally consonant with the data.

49.6 In 47.11 we defined functions

$$a(\alpha) = \frac{1}{\sqrt{(n\pi)}} \sum_{t=1}^n u_t \cos \alpha t, \quad (49.8)$$

$$b(\alpha) = \frac{1}{\sqrt{(n\pi)}} \sum_{t=1}^n u_t \sin \alpha t. \quad (49.9)$$

We showed that the intensity $I(\alpha)$, defined as the sum of squares of $a(\alpha)$ and $b(\alpha)$, was equal in the limit to the spectral density function $w(\alpha)$ multiplied by σ^2/π . We graphed $w(\alpha)$ for a Markoff and a Yule scheme in Fig. 47.4 and 47.5. A few practical examples are given in Fig. 49.1 to 49.3 for comparison with the correlograms of Exercises 47.20

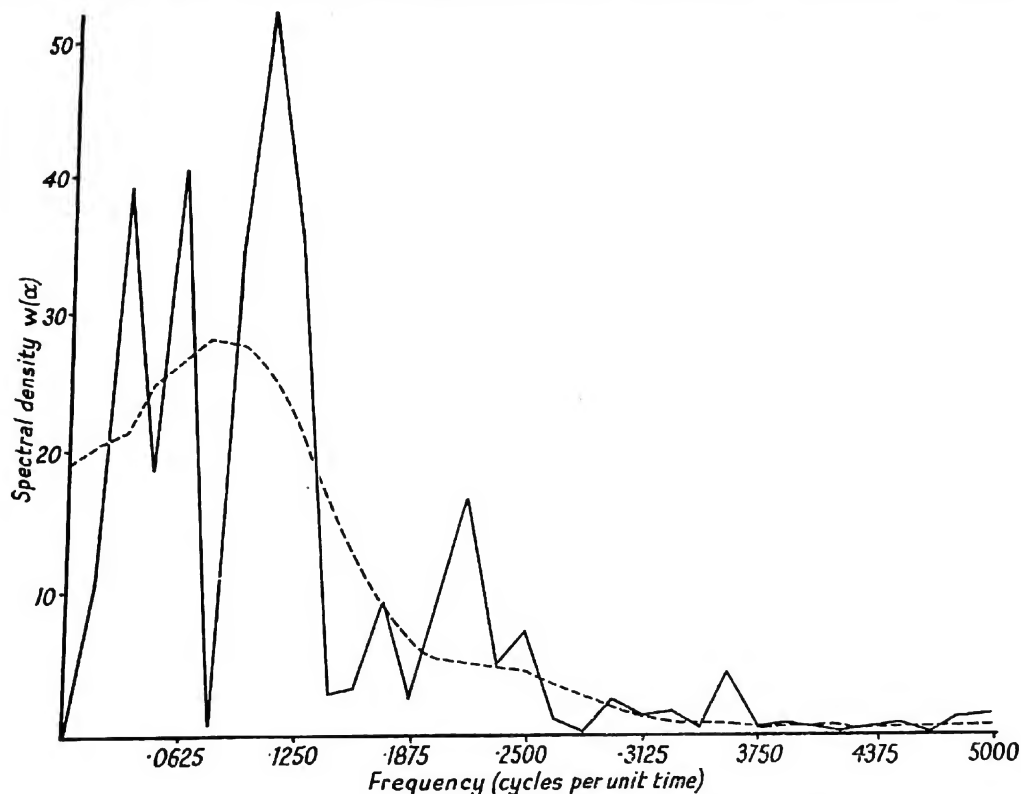


Fig. 49.3—Power spectrum of the data of Table 47.4 (second-order autoregressive scheme)

The dotted line is a smoothed spectrum using a Parzen window (Exercise 49.7) and the first fifteen covariances.

to 47.22 (namely the Beveridge wheat series, the marriage-rate data of Table 47.2, and the artificial Yule scheme of Table 47.4).

Observational material often presents these wild fluctuations and we shall see presently why this is so.

49.7 It is sometimes convenient to take as ordinate the logarithm of $w(\alpha)$ rather than $w(\alpha)$ itself. This avoids over-emphasis of the larger intensities and also has the advantage, as we shall see later, that the error bands in certain classes of estimation

are of constant width (cf. 49.15). For the most part we use $w(\alpha)$, but some examples in the next chapter are based on $\log w(\alpha)$.

The sums $a(\alpha)$ and $b(\alpha)$, being weighted sums of variables with the same variance, will be close to normality for stationary series. We will first consider the behaviour of the spectrum when harmonic or trend terms are present.

49.8 Suppose that the series u_t consists of a harmonic term with angular frequency α added to other terms which are not correlated with it:

$$u_t = c \sin \alpha t + g. \quad (49.10)$$

We calculate

$$\sum_{t=1}^n \sin \alpha t e^{i\beta t}$$

and find that it is equal to

$$\frac{1}{\sin \{\frac{1}{2}(\alpha - \beta)\}} [\cos \{(n + \frac{1}{2})(\alpha - \beta)\} - \cos \{\frac{1}{2}(\alpha - \beta)\} - i \sin \{(n + \frac{1}{2})(\alpha - \beta)\} + i \sin \{\frac{1}{2}(\alpha - \beta)\}]$$

+ a similar term with $+\beta$ in place of $-\beta$. (49.11)

In the neighbourhood of $\beta = \alpha$ this is dominated by the first term. The sum $\sum g e^{i\beta t}$ is, by hypothesis, of negligible size. Hence the intensity $I(\alpha)$ is the sum of squares of real and imaginary parts in (49.11), multiplied by $c^2/n\pi$. We find

$$I(\beta) = \frac{c^2 \sin^2 \{\frac{1}{2}n(\alpha - \beta)\}}{4\pi n \sin^2 \{\frac{1}{2}(\alpha - \beta)\}}. \quad (49.12)$$

The corresponding periodogram ordinate is, from (47.30),

$$S^2(\beta) = \frac{c^2 \sin^2 \{\frac{1}{2}n(\alpha - \beta)\}}{n^2 \sin^2 \{\frac{1}{2}(\alpha - \beta)\}}. \quad (49.13)$$

Now suppose that

$$\alpha - \beta = \frac{2\pi m}{n}, \quad n \text{ large, } m \text{ finite.} \quad (49.14)$$

We have then, to a close approximation,

$$S(\beta) = \frac{c^2 \sin^2 m\pi}{(m\pi)^2}. \quad (49.15)$$

Hence at $\beta = \alpha$ the periodogram will have a peak of amplitude c^2 and this will be flanked on either side by lesser peaks of diminishing intensity at distance $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}$, etc., from it.

A similar effect appears in the power spectrum, except that at $\beta = \alpha$ the ordinate becomes infinite in theory and may be very large in practice. The reason for choosing the divisor $\sqrt{(n\pi)}$ in (49.8) and (49.9) rests on the fact that we wish the ordinate to give the value of the power spectrum or an estimate of it. If u is a purely random series, in the limit

$$I(\alpha) = \sigma^2/\pi, \quad (49.16)$$

so that the ordinate of the spectrum is the same for all frequencies. In the periodogram it would, theoretically, be zero.

49.9 Whether the "side-bands" given by (49.15) show up either in spectrum or periodogram depends to some extent on the intervals of frequency at which I or S are computed. If the periodogram is plotted with period as abscissa (as, in our definition, it is) the side-bands become wider for increasing period. In fact, let

$$\lambda = \frac{2\pi}{\alpha}, \quad \mu = \frac{2\pi}{\beta}. \quad (49.17)$$

Then from (49.14)

$$\frac{1}{\lambda} - \frac{1}{\mu} = \frac{m}{n}$$

or approximately

$$\mu - \lambda = \frac{m\lambda^2}{n}, \quad (49.18)$$

so that the width of the side-band peaks depends on λ .

Fig. 49.4 gives the periodogram of the Beveridge series for comparison with Fig. 49.1. The values were calculated by Beveridge, at first on a grid of wavelengths of fairly equal width, but supplemented by additional values where peaks seemed to be indicated.

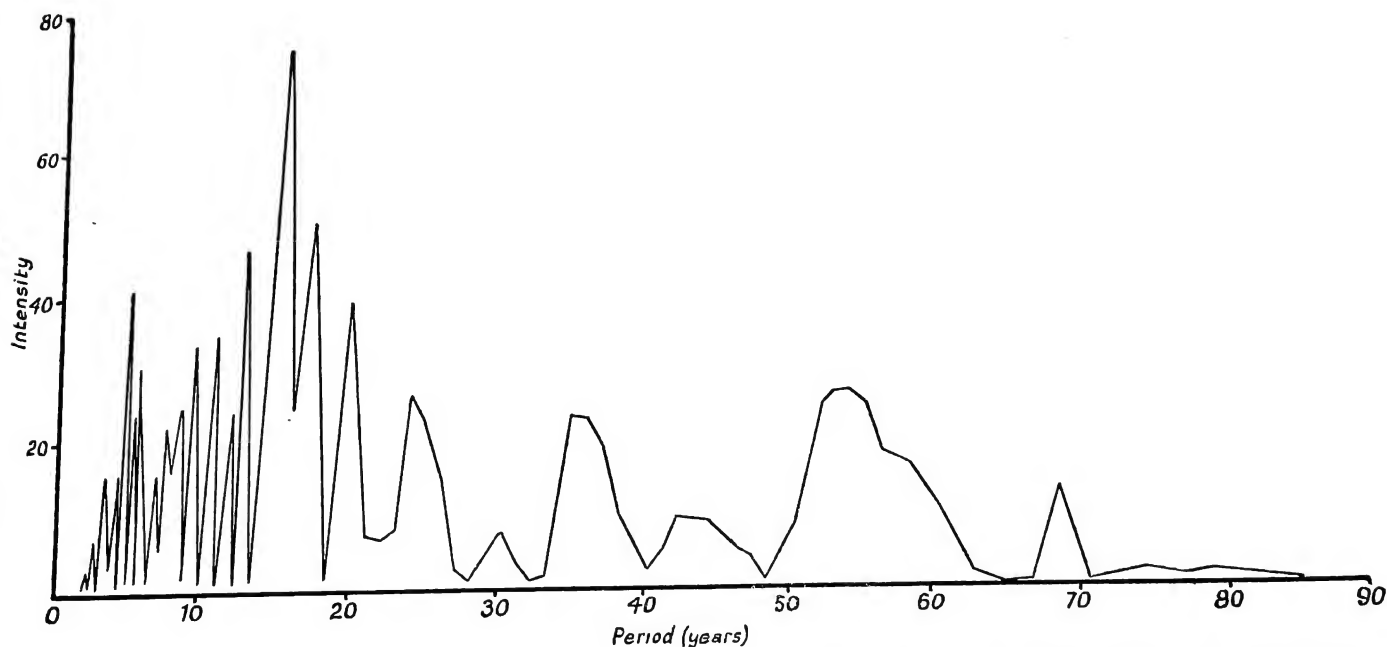


Fig. 49.4—Periodogram of the Beveridge wheat-price index series, for comparison with the power spectrum of Fig. 49.1

Example 49.1

It sometimes avoids tedious summation, and makes the essential point for asymptotic results, if we replace sums by integrals. For example, with n large,

$$\sum_{t=1}^n \sin \alpha t \sin \beta t$$

may be replaced by

$$\int_0^T \sin \alpha t \sin \beta t dt,$$

where T is the length of the series. The integral is seen to be

$$\frac{1}{2} \left[\frac{\sin \{(\alpha - \beta)T\}}{\alpha - \beta} - \frac{\sin (\alpha + \beta)T}{\alpha + \beta} \right].$$

Likewise, to the same degree of approximation

$$\sum_{t=1}^n \sin \alpha t \cos \beta t \, dt = \frac{1}{2} \left[-\frac{\cos \{(\alpha - \beta)T\} - 1}{\alpha - \beta} - \frac{\cos \{(\alpha + \beta)T\} - 1}{\alpha + \beta} \right].$$

The intensity near $\alpha = \beta$ is then given by

$$I(\beta) = \frac{\sin^2 \left\{ \frac{1}{2}(\alpha - \beta)T \right\}}{\pi T(\alpha - \beta)^2} \quad (49.19)$$

and the limiting case when $\alpha - \beta$ tends to zero may be discussed as before.

Non-harmonic periodicities

49.10 It must be remembered that a peak in the spectrum, interpreted as a harmonic, is only unrelated to other peaks if they all relate to pure sine or cosine terms. If there is present a periodic term which is not a simple harmonic there may be several peaks in the spectrum corresponding to it.

Consider a somewhat extreme case in which the periodicity is of the type shown in Fig. 49.5.

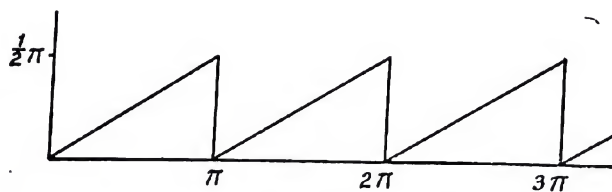


Fig. 49.5 (see text)

This is, in fact, the graph of $\frac{1}{2}x$ in the range $0 \leq x < \pi$, continually repeated. Now we have the Fourier expansion

$$\frac{1}{2}x = \sin x - \frac{1}{2} \sin 2x + \frac{1}{3} \sin 3x - \dots, \quad 0 \leq x < \pi. \quad (49.20)$$

Thus, in the spectrum there will be peak intensities at frequencies 1, 2, 3, etc. In the periodogram the intensities would form a series with diminishing amplitudes proportional to $1, \frac{1}{4}, \frac{1}{9}$, etc. For non-harmonic periodic elements, therefore, there is always the possibility of the fundamental frequency being echoed along the spectrum.

Example 49.2

Let us consider what happens if the series has a linear trend in it. In fact, let us take a pure trend $u_t = t$ and apply spectrum analysis to it. Approximately, as in Example 49.1, we have

$$\begin{aligned} \int_0^T t \sin \alpha t \, dt &= \left[-\frac{t \cos \alpha t}{\alpha} \right]_0^T + \int_0^T \frac{\cos \alpha t}{\alpha} \, dt \\ &= -\frac{T \cos \alpha T}{\alpha} + \frac{\sin \alpha T}{\alpha^2}. \end{aligned} \quad (49.21)$$

Likewise,

$$\int_0^T t \cos \alpha t \, dt = \frac{T \sin \alpha T}{\alpha} + \frac{\cos \alpha T - 1}{\alpha^2}. \quad (49.22)$$

Thus the intensity is given by

$$\begin{aligned} I(\alpha) &= \frac{1}{\pi T} \left\{ \frac{T^2}{\alpha^2} + O(T) \right\} \\ &= \frac{T}{\pi \alpha^2} + O(1). \end{aligned} \quad (49.23)$$

The power spectrum (T constant) would therefore be a curve of type $y = 1/x^2$, with large intensity at the origin. The periodogram, on the other hand, would have

$$S^2(\alpha) = \frac{4}{\alpha^2} = \frac{\lambda^2}{\pi^2}, \quad \text{where } \lambda \text{ is the wavelength.} \quad (49.24)$$

The results are understandable in general terms. A trend is like a long wave which is equivalent to a low frequency. Evidently, if low frequencies are of interest, every endeavour must be made to remove trend from a series before spectrum methods are applied.

Test for the spectral ordinate

49.11 Harmonic terms in a series may be likened to point-densities in a probability distribution; in the spectrum they define lines, not continuous densities, although, of course, in practice these lines are blurred for finite series. We proceed to consider the behaviour of the spectrum for stationary series of the non-deterministic type, which can be represented as the weighted sum (finite or infinite) of a series of random variables.

Consider first of all the sums $a(\alpha)$ and $b(\alpha)$ of (49.8) and (49.9) when u_t is a random series with zero autocorrelations and variance σ^2 . Since, for large n ,

$$\frac{1}{n} \sum_{k=1}^n \cos^2 \alpha k \rightarrow \frac{1}{2} \quad \frac{1}{n} \sum_{k=1}^n \sin^2 \alpha k \rightarrow \frac{1}{2}, \quad (49.25)$$

$$\frac{1}{n} \sum_{k=1}^n \cos \alpha k \sum_{k=1}^n \sin \alpha k \rightarrow 0, \quad (49.26)$$

we see that a, b are independent $N(0, \sigma^2/(2\pi))$ variables. Hence $2\pi I/\sigma^2 = 2\pi(a^2 + b^2)/\sigma^2$ is distributed as χ^2 with two degrees of freedom. Equivalently, the sum S^2 in the periodogram is distributed as

$$dF = \frac{n}{4\sigma^2} \exp\left(-\frac{nS^2}{4\sigma^2}\right) dS^2. \quad (49.27)$$

It follows that for the spectral ordinate, asymptotically,

$$E(I) = \frac{\sigma^2}{\pi}, \quad E(w) = 1, \quad (49.28)$$

$$\text{var } I = \frac{\sigma^4}{\pi^2}, \quad \text{var } w = 1 = \{E(w)\}^2. \quad (49.29)$$

Thus for a random series the standard error of the spectral ordinate is of the same order of magnitude as the ordinate itself.

The distribution (49.27) has been used to provide a test for ordinates in the periodogram. The probability that S^2 exceeds some value $4\sigma^2\kappa/n$ is $e^{-\kappa}$. In 1914, G. Walker pointed out that if $e^{-\kappa}$ is small, the probability that m independent ordinates should not exceed $4\sigma^2\kappa/n$ is $(1 - e^{-\kappa})^m$, so the chance that one at least exceeds that amount is

$$1 - (1 - e^{-\kappa})^m.$$

Davis (1941) tabulated this function. Fisher (1929), remarking that the test depended on σ^2 , effectively Studentized it. Davis also tabulated the function which emerges from the analysis.

49.12 It is very remarkable that the result of equation (49.29) is, for large samples, generally true of stationary series of the non-deterministic type, a result which is essentially due to Bartlett (see, e.g., his book of 1955). It will be convenient to summarize (49.8) and (49.9) in a single formula

$$J(\alpha) = a(\alpha) + ib(\alpha) = \frac{1}{\sqrt{(n\pi)}} \sum_{t=1}^n u_t e^{i\alpha t}. \quad (49.30)$$

The expectation of $J(\alpha)$ is zero. If u_t is a random series with zero autocorrelations and variance σ^2 , we have

$$\begin{aligned} E\{J(\alpha)J(\beta)\} &= \frac{\sigma^2}{n\pi} \sum_{t=1}^n e^{i\alpha t + i\beta t} \\ &= \frac{\sigma^2}{n\pi} \frac{e^{i(\alpha+\beta)} \{1 - e^{in(\alpha+\beta)}\}}{1 - e^{i(\alpha+\beta)}}. \end{aligned} \quad (49.31)$$

If α, β are of the form $2\pi p/n$, p integral, this vanishes. Similarly it follows that $J(\alpha)$ is uncorrelated with the complementary $J^*(\alpha) = a(\alpha) - ib(\alpha)$. Thus $a(\alpha)$ and $b(\alpha)$ are uncorrelated in this case and the corresponding spectral ordinates are uncorrelated.

We have $I(\alpha) = J(\alpha)J^*(\alpha)$, and putting $\beta = -\alpha$ in (49.31) we confirm the result that

$$E\{I(\alpha)\} = \frac{\sigma^2}{\pi}.$$

We have further

$$\begin{aligned} E\{I(\alpha)I(\beta)\} &= \frac{1}{n^2\pi^2} E\left\{ \sum_{t=1}^n u_t e^{i\alpha t} \sum_{s=1}^n u_s e^{-i\alpha s} \sum_{k=1}^n u_k e^{i\beta k} \sum_{l=1}^n u_l e^{-i\beta l} \right\} \\ &= \frac{1}{n^2\pi^2} \sum E(u_t u_s u_k u_l) \exp\{i(\alpha t - \alpha s + \beta k - \beta l)\}. \end{aligned} \quad (49.32)$$

The expectations vanish unless $t = s = k = l$ (giving the fourth-order moment of u) or the suffixes are equal in pairs. If $t = s$ and $k = l$ the corresponding term is $E\{I(\alpha)\}E\{I(\beta)\}$. Hence we find

$$\text{cov}\{I(\alpha), I(\beta)\} = \frac{\kappa_4}{n\pi^2} + \frac{\sigma^4}{n^2\pi^2} \left\{ \frac{1 - \cos\{n(\alpha+\beta)\}}{1 - \cos(\alpha+\beta)} + \frac{1 - \cos\{n(\alpha-\beta)\}}{1 - \cos(\alpha-\beta)} \right\}. \quad (49.33)$$

If $\alpha = \beta$ we find, since $1 - \cos \theta = \frac{1}{2}\theta^2$ for θ small,

$$\text{var } I(\alpha) = \frac{\sigma^4}{\pi^2} + O(n^{-1}), \quad (49.34)$$

confirming (49.29).

If u is non-normal, the covariance of $I(\alpha)$ and $I(\beta)$ is of order $1/n$. If u is normal it is of order $1/n^2$ and further is zero if α, β are of the form $2\pi p/n$, p integral.

49.13 Consider now the case when u_t is a weighted average of random variables ε , say

$$u_t = \sum_0^\infty g_s \varepsilon_{t-s}. \quad (49.35)$$

We have

$$\begin{aligned}
 J_u(\alpha) &= \frac{1}{\sqrt{n\pi}} \sum_{s=0}^{\infty} \sum_{t=1}^n g_s \varepsilon_{t-s} e^{i\alpha t} \\
 &= \frac{1}{\sqrt{n\pi}} \sum_{s=0}^{\infty} \sum_{t=1}^n \varepsilon_{t-s} e^{i\alpha(t-s)} e^{i\alpha s} g_s \\
 &= \frac{1}{\sqrt{n\pi}} \sum_{k=0}^{\infty} \varepsilon_{t-k} e^{i\alpha k} \sum_{s=0}^{\infty} e^{i\alpha s} g_s, \text{ approximately} \\
 &= J_\varepsilon(\alpha) h(\alpha)
 \end{aligned} \tag{49.36}$$

where $h(\alpha)$ is the transform of g_s , namely

$$h(\alpha) = \sum_{s=0}^{\infty} g_s e^{i\alpha s}. \tag{49.37}$$

We have at once

$$I_u(\alpha) = I_\varepsilon(\alpha) h(\alpha) h^*(\alpha), \tag{49.38}$$

which is another form of the result obtained for the effect of a transfer function in 47.24 in the context of a continuous series. Further we obtain

$$E\{I_u(\alpha)\} = h(\alpha) h^*(\alpha) E\{I_\varepsilon(\alpha)\} \tag{49.39}$$

and asymptotically,

$$\text{var } I_u(\alpha) = [E\{I_u(\alpha)\}]^2. \tag{49.40}$$

Smoothing the spectrum

49.14 These results provide us with a novel problem in estimation. The observed ordinate in the spectrum for a series of length n does not have a variance of order $1/n$, but of order w^2 . Furthermore, since the ordinates for values of α equal to $2\pi p/n$ are uncorrelated (exactly for normal variation and approximately otherwise), ordinates calculated for such values are effectively independent. The observed spectrum will thus fluctuate violently—Fig. 49.1 is a good example—and is a most unreliable estimator of the parent spectrum.

We shall attempt to overcome this difficulty by smoothing the spectrum, replacing $I(\alpha)$ by a weighted sum of neighbouring ordinates. This will render the estimator well-behaved in the sense of having a small variance, but to obtain such a result we have to pay a price in the form of bias in the estimator itself.

49.15 Let us take a weighting function $h(u)$ obeying the conditions

$$h(u) = h(u + 2\pi), \tag{49.41}$$

$$\int_{-\pi}^{\pi} h(u) du = 1. \tag{49.42}$$

This function is variously known as a “kernel” or “spectral window.” If $I(\alpha)$ is the estimated intensity we construct the smoothed function

$$\begin{aligned}
 I_A(\alpha) &= \int_{-\pi}^{\pi} h(u) I(\alpha - u) du \\
 &= \int_{-\pi}^{\pi} h(\alpha - u) I(u) du.
 \end{aligned} \tag{49.43}$$

If $I(\alpha)$ is unbiased we see, on taking expectations, that $I_A(\alpha)$ will, in general, be

biased, being a weighted average. To reduce the bias we desire $h(u)$ to be concentrated in a narrow range in the neighbourhood of $u = 0$, in which case the integral on the right in (49.43) will give an *approximately* unbiased result. Unless $h(u)$ is, trivially, the unit function at $u = 0$ there will, however, be some loss of resolution. A highly concentrated $h(u)$ may be thought of as possessing an effective range which is much narrower than the full range of definition $-\pi$ to $+\pi$, and this effective range is sometimes known as the "bandwidth" of the spectral window.

We may approximate to the integral (49.43) by a sum

$$I_A(\alpha) = \frac{2\pi}{n} \sum h(u_j) I(\alpha - u_j), \quad u_j = 2\pi j/n. \quad (49.44)$$

Since the values of I are independent we then have

$$\text{var } I_A(\alpha) = \frac{4\pi^2}{n^2} \sum h^2(u_j) \text{var } I(\alpha - u_j)$$

and using (49.40), this is approximately

$$\begin{aligned} &\doteq \frac{4\pi^2}{n^2} \sum h^2(u_j) I^2(\alpha - u_j) \\ &\doteq \frac{2\pi}{n} \int_{-\pi}^{\pi} h^2(u) I^2(\alpha - u) du. \end{aligned} \quad (49.45)$$

If $h(u)$ is concentrated in a narrow bandwidth this will give us, approximately,

$$\text{var } I_A(\alpha) = \frac{2\pi}{n} I^2(\alpha) \int_{-\pi}^{\pi} h^2(u) du. \quad (49.46)$$

Thus, provided that the integral is bounded, the variance is now of the order of $1/n$. It also follows that, to the same degree of approximation, $\text{var } \log I_A(\alpha)$ is a constant, and hence that $\log I_A$ has confidence intervals of constant width.

It may also be shown that the correlation between $I(\alpha)$ and $I(\beta)$ is approximately equal to

$$\frac{\int_{-\pi}^{\pi} h(u) h(u + \alpha - \beta) du}{\int_{-\pi}^{\pi} h^2(u) du}, \quad (49.47)$$

and as this is positive for any acceptable weighting function the use of the word "smoothing" is justified.

Calculation of spectra

49.16 For the calculation of spectral ordinates in practice we do not work out the sums (49.8) and (49.9) for varying values of α and then compute the intensity or the spectral density. In fact, we have

$$\begin{aligned} I(\alpha) &= \frac{1}{n\pi} \left\{ \left(\sum_1^n u_t \cos \alpha t \right)^2 + \left(\sum_1^n u_t \sin \alpha t \right)^2 \right\} \\ &= \frac{1}{n\pi} \sum_{s,t=-n}^n u_s u_t (\cos \alpha t \cos \alpha s + \sin \alpha t \sin \alpha s) \\ &= \frac{1}{\pi} \sum_{-n}^n c_k \cos k\alpha, \end{aligned} \quad (49.48)$$

where c_k is a covariance-type expression defined by

$$c_k = \frac{1}{n} \sum_{t=1}^{n-k} u_t u_{t+k}. \quad (49.49)$$

For infinite n this reduces to the known expression (cf. (47.27))

$$I(\alpha) = \frac{\sigma^2}{\pi} w(\alpha) = \frac{\sigma^2}{\pi} \sum_{-\infty}^{\infty} \rho_k \cos k\alpha. \quad (49.50)$$

Calculation of the spectrum usually proceeds from (49.48). In any case we cannot compute c_k for $k > n-1$ and in practice would rarely wish to go as far as $n-1$ serial correlations. Let us then consider the estimator

$$I_q(\alpha) = \frac{1}{\pi} \sum_{-q}^q \lambda_k c_k \cos k\alpha, \quad (49.51)$$

based on q serial correlations. (The λ 's are constants to be chosen at convenience for the purpose of improving the estimator.) This is equivalent, in parental form, to

$$I_q(\alpha) = \frac{\sigma^2}{\pi} \sum_{-q}^q \lambda_k \rho_k \cos k\alpha$$

or

$$w_q(\alpha) = \sum_{-q}^q \lambda_k \rho_k \cos k\alpha. \quad (49.52)$$

But from (47.22)

$$\rho_k = \frac{1}{\pi} \int_0^\pi w(u) \cos ku \, du,$$

and hence on substitution in (49.52) we find

$$\begin{aligned} w_q(\alpha) &= \frac{1}{\pi} \sum_{-q}^q \lambda_k \int_0^\pi w(u) \cos ku \cos k\alpha \, du \\ &= \int_{-\pi}^\pi w(u) \left\{ \frac{1}{2\pi} \sum_{-q}^q \lambda_k \cos ku \cos k\alpha \right\} du. \end{aligned} \quad (49.53)$$

The use of (49.51) is then, asymptotically, equivalent to smoothing the spectrum by the weighting function

$$h(\beta) = \frac{1}{2\pi} \sum_{-q}^q \lambda_k \cos k\beta \cos k\alpha. \quad (49.54)$$

Provided that $\lambda_0 = 1$, this obeys conditions (49.41) and (49.42).

We also have

$$\int_{-\pi}^\pi h^2(\beta) \, d\beta = \frac{1}{2\pi} \sum_{-q}^q \lambda_k^2 \cos^2 k\alpha. \quad (49.55)$$

Example 49.3

Suppose, in the first instance, we take all λ 's equal to unity. Then

$$\begin{aligned} h(\beta) &= \frac{1}{2\pi} \sum_{-q}^q \cos k\beta \cos k\alpha \\ &= \frac{1}{\pi} \sum_{-q}^q [\cos \{(\alpha + \beta)k\} + \cos \{(\alpha - \beta)k\}] \\ &= \frac{1}{\pi} \left[\frac{\sin \{(q + \frac{1}{2})(\alpha + \beta)\}}{\sin \{\frac{1}{2}(\alpha + \beta)\}} + \frac{\sin \{(q + \frac{1}{2})(\alpha - \beta)\}}{\sin \{\frac{1}{2}(\alpha - \beta)\}} \right], \end{aligned} \quad (49.56)$$

with

$$\int_{-\pi}^{\pi} h^2(\beta) d\beta = \frac{1}{2\pi} \left[q + \frac{1}{2} + \frac{\sin \{(2q+1)\alpha\}}{2 \sin \alpha} \right]. \quad (49.57)$$

Example 49.4 (Bartlett, 1950)

Take

$$\lambda_k = 1 - \frac{k}{q}. \quad (49.58)$$

We find

$$h(\beta) = \frac{1}{\pi} \left\{ \frac{\sin^2 \{\frac{1}{2}q(\alpha+\beta)\}}{q \sin^2 \{\frac{1}{2}(\alpha+\beta)\}} + \frac{\sin^2 \{\frac{1}{2}q(\alpha-\beta)\}}{q \sin^2 \{\frac{1}{2}(\alpha-\beta)\}} \right\} \quad (49.59)$$

and

$$\int_{-\pi}^{\pi} h^2(\beta) d\beta = \frac{1}{2\pi} \left\{ \frac{(q-1)(2q-1)}{6q} + \frac{1}{2q \sin^2 \alpha} - \frac{\sin 2q\alpha \cos \alpha}{4q^2 \sin^3 \alpha} \right\}. \quad (49.60)$$

Example 49.5 (Daniell, 1946)

Take

$$\lambda_k = \frac{\sin kh}{kh}, \quad h > 0. \quad (49.61)$$

We have the known integrals

$$\left. \begin{aligned} \int_0^{\infty} \frac{\sin px \cos qx}{x} dx &= \frac{1}{2}\pi, & |p| > |q| \\ &= \frac{1}{4}\pi, & |p| = |q| \\ &= 0, & |p| < |q| \end{aligned} \right\} \quad (49.62)$$

The weighting function is then given by

$$h(\beta) = \frac{1}{2\pi} \left\{ 1 + 2 \sum_{k=1}^q \frac{\cos \beta k \cos \alpha k \sin hk}{hk} \right\}$$

which is approximated by the integral

$$\left. \begin{aligned} \frac{1}{2\pi} \int_0^{\infty} \frac{\sin hx}{hx} [\cos \{\frac{1}{2}(\alpha+\beta)x\} + \cos \{\frac{1}{2}(\alpha-\beta)x\}] dx \\ = \frac{1}{2h}, & \quad h \geq \alpha - \beta \geq -h, \\ = 0 & \quad \text{elsewhere.} \end{aligned} \right\} \quad (49.63)$$

Various other kernels have been suggested, notably by Blackman and Tukey (1958) and Parzen (1961). See Exercises 49.5–7 and a review by Jenkins (1961).

Estimation of spectral densities

49.17 The problem of estimating spectral densities has received a great deal of attention, and a complete account of the subject (which is itself by no means complete) would occupy more space than we can allot to it. We must content ourselves with a summary account of the principles.

The object of the estimation is to provide good estimates of the ordinates along the length of the power spectrum. (This is not usually the ultimate object of the analysis, a point which is apt to be overlooked.) We have seen that the ideal may be unattainable

for various reasons. We therefore introduce the kernel or spectral window to smooth out the grosser irregularities in the observed spectrum. A "good" kernel will be relatively narrow in range, but no kernel can be perfect, and its values may be unduly influenced by casual peaks in the spectrum—there is then said to be "leakage" round the edges of the "spectral window" which we are using to scrutinize part of the spectrum. We are thus led to consider the effectiveness of different kernels in smoothing, and hence introducing reliability, as against averaging, and hence introducing bias. For some of the procedures possible in this context reference may be made to Whittle (1957), who considers a prior distribution of spectral ordinates, Blackman and Tukey (1958), who discuss the use of prior analysis of the data, and Parzen (1961), who considers, among other things, the prior determination of the rate of decay of the autocorrelations.

49.18 Leakage causes trouble, especially in estimates of the low part of the spectrum, for the kernel, though itself small in the outlying parts of its range, may swamp the average when multiplied by a high value of the spectral ordinate. For this reason Blackman and Tukey (1958) introduced a process known as *pre-whitening*, the object of which is to filter the series so that the peaks are flattened out. For example, if the original spectrum has a peak at α_1 and we can transform the original series so that this peak is flattened out, the estimate at some other point α_2 will no longer be distorted by α_1 . This is obviously a rather dangerous procedure, but fortunately we can afterwards *recolour* the spectrum, in the terminology of Nerlove (1964). The basic idea rests on the result we proved in 47.24, that if $v(t)$ is a filtered series derived from $u(t)$ by a linear filter, then

$$w_v(\alpha) = w_u(\alpha) |f(\alpha)|^2, \quad (49.64)$$

where $f(\alpha)$ is the transfer function of the filter itself. Knowing the filter, we can always recover the spectrum of the original series from the estimated spectrum of the transformed series. The procedure has been examined by Hext (1964). It may well be desirable to use different procedures for different parts of the spectrum.

49.19 Daniels (1962) develops some alternative approaches. He considers first of all a preliminary smoothing by a kernel chosen so as to obey a given criterion, e.g. so as to achieve the minimum tolerable resolution. Then he *unsmooths* the spectrum by setting up a routine which improves the resolution at the expense of the sampling variance until no further useful change is detectable in the fitted spectrum. Two unsmoothing processes are discussed, one approximating the spectrum locally by a polynomial, the other based on differences of the spectrum. The process is empirical in the sense that it uses the data to determine the estimator, and it requires a good deal of computation, but it at least proceeds by successive approximation to a stable solution.

49.20 In connexion with equation (49.64) we may note some work by Hannan (1960) and Durbin (1961) concerning the effect of seasonal variation and trend-elimination on the spectrum. We have already remarked on the problems created by the elimination of trend in distorting the residuals. With (49.64) we can regenerate the

spectrum of the residuals undistorted by trend-removal, at least in theory. But this does not, of course, mean that we can regenerate the residuals themselves.

49.21 It must always be remembered that we are not interested in the periodogram or the power spectrum for its own sake, except perhaps in the domain of physics or electrical engineering, where an ordinate in the spectrum can be given a physical interpretation as the amount of power which is obtainable at a given cycle frequency. For general statistical purposes the spectrum is a diagnostic instrument whose main use is to suggest an appropriate model to generate the series under observation. Interest therefore tends to be focussed on the testing of hypotheses concerning the model, rather than testing particular ordinates in a correlogram or a spectrum, and this is a subject we consider in the next chapter. For more extensive studies of spectrum analysis we may refer to the books by Blackman and Tukey (1958), Grenander and Rosenblatt (1957), and Granger and Hatanaka (1964), and the symposium edited by Rosenblatt (1963).

Unequal time-intervals

49.22 Finally we may add a few comments on a point of some practical importance where daily or monthly observations are concerned. Suppose we have a series u_1, u_2 , etc. observed at intervals of mt , so that our information consists of observations $u(m)$, $u(2m)$, etc. For example, we may observe a daily series once a month, in which case m is, on the average, about 30. Suppose further that the intervals between observations now vary about m to some extent, so that we have instead observations $u(m + \varepsilon_1)$, $u(2m + \varepsilon_2) \dots$, etc. If u has zero mean and unit variance, which we may assume without loss of generality, the autocorrelations of the original observations are given by

$$\rho(km) = E[u(tm)u\{(t+k)m\}], \quad k = 0, 1, \dots \quad (49.65)$$

Those of the second series are, say $\rho^*(km)$, given by

$$\begin{aligned} \rho^*(km) &= E \frac{1}{n} \sum_{p=1}^n u(pm + \varepsilon_p) u\{(p+k)m + \varepsilon_{p+k}\} \\ &= \frac{1}{n} \sum_{p=1}^n \rho(km + \varepsilon_p - \varepsilon_{p+k}). \end{aligned} \quad (49.66)$$

Expanding ρ either as a Taylor series or an equivalent series of differences, we then have, to the second order in ε ,

$$\begin{aligned} \rho^*(km) &= \rho(km) + \frac{1}{n} \sum (\varepsilon_p - \varepsilon_{p+k}) \rho'(km) \\ &\quad + \frac{1}{2n} \sum (\varepsilon_p - \varepsilon_{p+k})^2 \rho''(km). \end{aligned} \quad (49.67)$$

If ε has zero mean the second term is small and vanishes in the limit. Writing σ^2 for the variance of ε and τ_k for its k th autocorrelation, we then have

$$\rho^*(km) \doteq \rho(km) + \sigma^2 (1 - \tau_k) \rho''(km). \quad (49.68)$$

On the average, then, the autocorrelations are not seriously disturbed so long as σ^2 is small.

For example, consider observations made on the first of the month instead of daily. The average month-length, taking account of leap years, is 30.437 with a variance of 0.70. The first autocorrelation τ_1 is -0.42 , and the second is positive. On our scale $\sigma^2 = 0.70/(30.437)^2 = 0.00376$. The autocorrelations based on the first of the month will then, from (49.68), be only slightly affected on average, provided that ρ'' , or the second difference of ρ , is not very large, which is so.

Similar arguments apply to the power spectrum. Low frequencies are emphasized slightly, but the effect is negligible.

49.23 The matter stands differently for series which are aggregated, such as rainfall. The effect of differing time-intervals may then be serious, as is fairly evident when we remember that we may be comparing sums based (in the case of months and days) on 28 or 31 observations. It is, in our opinion, essential in such cases to standardize the data by reducing them to a period of constant length. This is particularly true for such data as output per working week or inputs per working month.

Granger (1963) has discussed the matter in more detail from the spectrum viewpoint. See also Quenouille (1958).

EXERCISES

49.1 $P(t)$ is a polynomial of degree k for $0 \leq t \leq T$. Show that asymptotically the ordinate in the periodogram corresponding to frequency α is given by

$$\frac{4P^2(T)}{\alpha^2 T^2} + O(T^{2k-s}).$$

49.2 A series has the value e^{ct} in the range 0 to T , $c > 0$. Show that asymptotically the ordinate in the periodogram corresponding to frequency α is

$$\frac{16 \exp(2cT)}{T^2} \sin^2 \frac{1}{2} \alpha T.$$

49.3 Given that

$$x = \frac{4}{\pi} \left(\sin x - \frac{\sin 3x}{3^2} + \frac{\sin 5x}{5^2} - \dots \right), \quad -\frac{1}{2}\pi \leq x \leq \frac{1}{2}\pi,$$

graph the series whose term is x over the range 0 to 4π . Compare with **49.10** and comment on the effects on the power spectrum.

49.4 Establish equation (49.47).

49.5 In (49.54) take

$$\lambda_k = 1 - 2a + 2a \cos(\pi k/q).$$

Show that

$$h = \frac{1}{\pi} \left[(1-2a) \frac{\sin(q+\frac{1}{2})\gamma}{\sin \frac{1}{2}\gamma} + a \left\{ \frac{\sin\{(q+\frac{1}{2})\gamma + \pi/q\}}{\sin(\frac{1}{2}\gamma + \pi/q)} + \frac{\sin\{(q+\frac{1}{2})\gamma - \pi/q\}}{\sin(\frac{1}{2}\gamma - \pi/q)} \right\} \right]$$

where $\gamma = \alpha - \beta$, together with a similar term obtained by putting $\gamma = \alpha + \beta$.

(Tukey, cf. Blackman and Tukey, 1958. They propose the values $a = 0.25$ or $a = 0.23$.)

THE ADVANCED THEORY OF STATISTICS

470

49.6 In a similar manner to the previous exercise, with

$$\lambda_k = 1 - k^2/q^2,$$

show that

$$h = \frac{1}{\pi q^2} \left\{ \frac{\frac{1}{2} \sin q\gamma \cos \frac{1}{2}\gamma}{\sin^3 \frac{1}{2}\gamma} - \frac{q \cos q\gamma}{\sin^2 \frac{1}{2}\gamma} \right\}.$$

(Parzen, 1961)

49.7 As in the previous exercise, with

$$\lambda_k = 1 - 6\left(\frac{k}{q}\right)^2 + 6\left(\frac{k}{q}\right)^3, \quad 0 \leq k \leq \frac{1}{2}q,$$

$$= 2\left(1 - \frac{k}{q}\right)^3, \quad \frac{1}{2}q \leq k \leq q,$$

show that

$$h = \frac{3}{4\pi q^3} \left\{ \frac{\sin \frac{1}{4}q\gamma}{\sin \frac{1}{4}\gamma} \right\}^4.$$

(Parzen, 1961)

49.8 If u_t is stationary and normal, and $\gamma_1 = \mu_3/\mu_2^{3/2}$, $\gamma_2 = \mu_4/\mu_2^2 - 3$, show that asymptotically γ_1 is $N(0, R_1)$ with

$$R_1^2 = 6 \sum_{-\infty}^{\infty} \rho_j^3$$

and that γ_2 is $N(0, R_2)$ with

$$R_2^2 = 24 \sum_{-\infty}^{\infty} \rho_j^4.$$

Show how this may be used to test for normality of a stationary process.

(Lomnicki, 1961)

49.9 The Buys-Ballot table. A series of $\rho\mu$ terms is written down in ρ rows of μ thus:

$$\begin{array}{cccc} \mu_1 & u_2 & \dots & u_\mu \\ u_{\mu+1} & u_{\mu+2} & \dots & u_{2\mu} \\ \cdot & \cdot & \cdot & \cdot \\ u_{(\rho-1)\mu+1} & u_{(\rho-1)\mu+2} & \dots & u_{\rho\mu} \end{array}$$

$$\text{Sums: } m_1 \quad m_2 \quad \dots \quad m_\mu$$

Show that the sums A and B entering into the periodogram are given by

$$A = \frac{2}{\rho\mu} \sum_{j=1}^{\mu} m_j \cos \frac{2\pi j}{\mu},$$

$$B = \frac{2}{\rho\mu} \sum_{j=1}^{\mu} m_j \sin \frac{2\pi j}{\mu}.$$

49.10 With reference to the previous exercise, consider

Show that if

$$\eta^2(\mu) = \text{var } m / \text{var } u.$$

where b_j is uncorrelated with periodic terms, then

$$\eta^2(\mu) = \left(\frac{a^2 \mu^2 \sin^2(n\pi/\lambda)}{2n^2 \sin^2(\mu\pi/\lambda)} + \frac{\mu}{n} \text{var } b \right) / \left(\frac{1}{2} a^2 + \text{var } b \right).$$

Hence show that, in the neighbourhood of λ , the graph of η as ordinate against μ as abscissa (Whittaker's periodogram) has a peak of breadth $2\lambda^2/n$, flanked by smaller peaks.
(Whittaker, 1911)

49.11 If an autoregressive series of Yule type (47.74) is subject to errors of observation which are independent from term to term, show that the serial correlations (except r_0) are reduced in constant proportion, say c . Hence, if α, β in

$$u_{t+2} + \alpha u_{t+1} + \beta u_t = \varepsilon_{t+2}$$

are estimated from the Yule-Walker equations (47.66) as a', b' , show that, for the series subject to error,

$$b'/a' > b/a,$$

where a, b refer to estimates for the same series not subject to error.

49.12 The following are the spectral densities computed for the series of Table 47.4 and Fig. 49.3 for smoothing with a Parzen window (Exercise 49.7) and various truncation points in the number q of serial correlations computed. Sketch the power spectra and note the disturbing effect of having q too large.

Spectral densities

Frequency (cycles per year)	No smoothing	$q = 15$	$q = 20$	$q = 30$
0	0.0000	19.3065	15.5732	11.6349
0.0156	10.4916	20.0120	18.2743	15.8810
0.0313	39.5419	21.8959	21.9328	24.0449
0.0469	18.6496	24.3610	25.1102	28.3488
0.0625	40.5257	26.6535	27.2150	26.1922
0.0781	1.1554	28.0290	29.1099	25.9544
0.0938	34.9011	27.8621	30.4033	33.5149
0.1094	52.4760	25.8092	28.9764	35.8784
0.1250	36.6309	22.0280	23.7428	25.2362
0.1406	2.7163	17.2623	16.5376	13.2443
0.1563	3.7282	12.6006	10.4166	7.8444
0.1719	9.6669	8.9806	6.9051	5.9028
0.1875	2.5688	6.7729	5.5586	4.6738
0.2031	0.0968	5.7301	5.3329	4.8532
0.2188	16.3082	5.2859	5.4431	5.9667
0.2344	5.5837	4.9314	5.3382	6.0461
0.2500	7.7572	4.4191	4.7363	4.9161
0.2656	1.3151	3.7474	3.7738	3.5804
0.2813	0.4748	3.0386	2.8106	2.5929
0.2969	2.4491	2.4260	2.0975	1.8886
0.3125	0.9332	1.9872	1.7045	1.3929
0.3281	1.9702	1.7203	1.5886	1.3985
0.3438	0.7401	1.5573	1.6005	1.7491
0.3594	4.0405	1.4114	1.5402	1.8142
0.3750	0.3481	1.2300	1.3106	1.3800
0.3906	0.4831	1.0181	0.9851	0.8537
0.4063	0.5796	0.8206	0.7045	0.5774
0.4219	0.1827	0.6847	0.5519	0.4695
0.4375	0.4529	0.6311	0.5288	0.4463
0.4531	0.4774	0.6486	0.5974	0.5478
0.4688	0.1512	0.7041	0.7097	0.7256
0.4844	1.1829	0.7582	0.8112	0.8687
0.5000	1.8403	0.7801	0.8521	0.9190

CHAPTER 50

TIME-SERIES: SOME FURTHER TOPICS

50.1 The theory of time-series has not reached a stage, and may never reach a stage, at which a clearly structured account of it can be given. To some extent this is due to the complicated nature of the subject—we have to take account not only of probability distributions but of their autocorrelations over time, and the embarrassing profusion of parameters which results may make it difficult to choose among a sizeable set of different hypotheses which are all consonant with the data. In some fields, especially in economics, experiences are rarely long enough to enable us to lean as heavily on our models as we can, for example, in physics. A run of fifty years' data is "long" as such series go, and even if longer, may arise from a system which is itself undergoing important structural change.

50.2 The advent of the electronic computer has removed most of the tedium which was a serious obstacle to former workers on time-series analysis, but there remain the problems of formulating and testing hypotheses or of setting up a model of the system under study. For this reason, a working statistician very often needs to call in aid a great deal of extraneous information of a non-statistical, perhaps a non-quantifiable kind, in order to define his problem and to set up his models. We shall not attempt a review of the considerations and methods to which he must have regard in this part of his work. We take them for granted, and in this final chapter shall consider the purely statistical aspects of the subject: estimation and hypothesis testing, multivariate extensions, and some related questions concerning identifiability and mixed regressive-autoregressive systems.

Estimation

50.3 We begin by emphasizing some general points which are peculiar to time-series analysis and, in one form or another, bedevil attempts to reach exact results in problems of estimation or hypothesis testing.

Consider the likelihood function of an autoregressive series. It will make for clarity if we discuss a Markoff scheme, although the argument is general. For a set of observations u_1, u_2, \dots, u_n we have

$$\left. \begin{aligned} u_1 &= \rho u_0 + \varepsilon_1 \\ u_2 &= \rho u_1 + \varepsilon_2 \\ &\vdots \\ u_n &= \rho u_{n-1} + \varepsilon_n \end{aligned} \right\} \quad (50.1)$$

If the probability distribution of the ε 's were known, say as $f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$, we might regard (50.1) as determining a variate transformation to new variables u_1, u_2, \dots, u_n . But here we encounter a difficulty in that u_0 is also involved. We have, in fact, n

variables ε and $n+1$ variables u . Let us then add to (50.1) the supplementary equation

$$u_0 = u_0. \quad (50.2)$$

We know that u_0 is dependent only on $\varepsilon_0, \varepsilon_{-1}$, etc. and hence is independent of ε_1 to ε_n . If its frequency function is $g(u_0)$ we then have for the joint distribution of $u_0, \varepsilon_1, \dots, \varepsilon_n$,

$$dF = f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n) g(u_0) d\varepsilon_1 d\varepsilon_2 \dots d\varepsilon_n du_0. \quad (50.3)$$

Let us now make the transformation to variables u_0, u_1, \dots, u_n . The Jacobian is easily seen to be unity, and we find

$$dF = f(u_1 - \rho u_0, u_2 - \rho u_1, \dots, u_n - \rho u_{n-1}) g(u_0) du_0 du_1 \dots du_n. \quad (50.4)$$

To manipulate this expression for the purpose of deriving estimators or tests we require to dismiss the element u_0 , which is unknown. (If it were known we should have started the series of observations with it.) There are several ways of doing this, but they all involve some sort of limitation on our inference:

- (a) we may assume u_0 known and make the inference conditional upon it;
- (b) we may make the sample circular, i.e. assume that $u_n = u_0$;
- (c) we may neglect u_0 by showing that asymptotically its effect is negligible.

50.4 The method we shall consider is the third. Suppose, for example, that the ε 's are distributed normally with unit variance. Then u_0 will be normal with variance $1/(1-\rho^2)$, and for $\log L$ we have, apart from constants,

$$\log L = \frac{1}{2} \log(1-\rho^2) - \frac{1}{2} \sum_{j=1}^n (u_j - \rho u_{j-1})^2 - \frac{1}{2} (1-\rho^2) u_0^2. \quad (50.5)$$

The term involving u_0 is seen to be $-\frac{1}{2} u_0^2 + \rho u_0 u_1$ and we integrate this out to obtain

$$\log L = \text{const.} + \frac{1}{2} \log(1-\rho^2) - \frac{1}{2} \sum_{j=2}^n (u_j - \rho u_{j-1})^2 - \frac{1}{2} (1-\rho^2) u_1^2. \quad (50.6)$$

For large n the summation dominates $\log L$, and asymptotically we have

$$\log L \sim -\frac{1}{2} \sum_{j=2}^n (u_j - \rho u_{j-1})^2. \quad (50.7)$$

We can estimate ρ by maximizing this likelihood, which is equivalent to minimization of a sum of squares over $(n-1)$ terms. Apart from the approximation, the results are what we would have got by treating the autoregression as an ordinary regression. The ML estimator is then

$$\hat{\rho} = \frac{\sum_2^n u_j u_{j-1}}{\sum_2^n u_{j-1}^2}.$$

50.5 The same point may be made in a different way. The variance of u_i is $1/(1-\rho^2)$ and the correlation of u_i and u_j is $\rho^{|i-j|}$. Thus the dispersion matrix of u_1, u_2, \dots, u_n is

$$\mathbf{V} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ . & . & . & . & . \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}. \quad (50.8)$$

The determinant is found to be $1/(1-\rho^2)$ and the inverse is

$$\mathbf{V}^{-1} = \begin{pmatrix} 1 & -\rho & 0 & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \dots & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}. \quad (50.9)$$

Hence, for normal variation, the log likelihood, apart from constants, is

$$\begin{aligned} \frac{1}{2} \log(1-\rho^2) - \frac{1}{2} \left\{ u_1^2 + \sum_{j=2}^{n-1} (1+\rho^2) u_j^2 + u_n^2 - 2\rho \sum_{j=2}^n u_j u_{j-1} \right\} \\ = \frac{1}{2} \log(1-\rho^2) - \frac{1}{2} \left\{ \sum_{j=2}^n (u_j - \rho u_{j-1})^2 + (1-\rho^2) u_1^2 \right\}, \end{aligned} \quad (50.10)$$

which brings us back to (50.6).

50.6 A second general point to notice concerns the relationship between autoregressive and moving-average schemes. We have already remarked in 47.18 that an autoregressive scheme is equivalent to a moving average of infinite extent. It might, then, be thought that a moving average, being actually of finite extent, would have simpler estimational properties. This turns out not to be so. We can illustrate the point by reference to the scheme

$$u_t = \varepsilon_t + \beta \varepsilon_{t-1}. \quad (50.11)$$

The dispersion matrix of the series (with unit variance for ε) is

$$\begin{pmatrix} 1+\beta^2 & \beta & 0 & 0 & \dots & 0 \\ \beta & 1+\beta^2 & \beta & 0 & \dots & 0 \\ 0 & \beta & 1+\beta^2 & \beta & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1+\beta^2 \end{pmatrix}. \quad (50.12)$$

With $\beta = -\rho$ this is nearly the same as (50.9), but the difference is not negligible and (50.12) is not so easy to invert. Consider, however, (50.8) with $\beta = -\rho$, namely

$$\frac{1}{1-\beta^2} \begin{pmatrix} 1 & -\beta & \beta^2 & \dots & (-\beta)^{n-1} \\ -\beta & 1 & -\beta & \dots & (-\beta)^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (-\beta)^{n-1} & (-\beta)^{n-2} & (-\beta)^{n-3} & \dots & 1 \end{pmatrix}. \quad (50.13)$$

If, in (50.11), we modify the model slightly so that ε_0 is zero, then $\text{var } u_1 = 1$. Likewise if ε_n has variance $1-\beta^2$, $\text{var } u_n = 1$. The other values are unaffected and hence (50.13) represents the inverse of the dispersion matrix of the modified scheme, which clearly is asymptotically the same as the scheme (50.11) since only two end terms have been altered.

Thus, to this degree of approximation, the log likelihood is given by

$$\begin{aligned} \log L = \text{const.} - \frac{1}{2} \log(1-\beta^2) - \frac{1}{2(1-\beta^2)} \left\{ \sum_{j=1}^n u_j^2 - 2\beta \sum_1^{n-1} u_j u_{j+1} \right. \\ \left. + 2\beta^2 \sum_1^{n-2} u_j u_{j+2} - \dots \right\}. \end{aligned} \quad (50.14)$$

In this expression all the observed serial covariances are involved, whereas for the autoregressive scheme we need only as many serial covariances as there are constants to estimate. Even if we neglect the terms outside braces in (50.14) we are still left with a cumbersome likelihood function to manage, and in particular the ML equations are intractable.

Example 50.1

It might be supposed that the difficulty could be overcome by using a different estimator. For example, we have for the first autocorrelation of (50.11)

$$\rho_1 = \beta/(1+\beta^2). \quad (50.15)$$

It is plausible, then, to estimate β by solving

$$\frac{b}{1+b^2} = r_1. \quad (50.16)$$

But unfortunately, as Whittle (1953a) showed, this is a very inefficient estimator.

In fact, for the asymptotic variance of r_1 (equation (48.9)) we have

$$n \text{ var } r_1 = \sum_{i=-\infty}^{\infty} \{\rho_i^2 + \rho_{i-1} \rho_{i+1} - 4\rho_1 \rho_i \rho_{i+1} + 2\rho_i^2 \rho_1^2\},$$

which, in our present case, reduces to

$$n \text{ var } r_1 = 1 - 3\left(\frac{\beta}{1+\beta^2}\right)^2 + 4\left(\frac{\beta}{1+\beta^2}\right)^4. \quad (50.17)$$

From (50.15) we have

$$dr_1 = \frac{1-b^2}{(1+b^2)^2} db, \quad (50.18)$$

and hence, asymptotically,

$$n \text{ var } b = \frac{1}{(1-\beta^2)^2} \{(1+\beta^2)^4 - 3\beta^2(1+\beta^2) + 4\beta^4\}. \quad (50.19)$$

For example, with $\beta = \frac{1}{2}$ we find

$$n \text{ var } b = \frac{389}{144}. \quad (50.20)$$

Taking $\log L$ in the form

$$\begin{aligned} & -\frac{1}{2(1-\beta^2)} \{\sum u_j^2 - 2\beta \sum u_j u_{j+1} + 2\beta^2 \sum u_j u_{j+2} - \dots\} \\ & = -\frac{2}{1-\beta^2} A, \text{ say,} \end{aligned}$$

we find

$$\begin{aligned} E(A) &= 1 - \beta^2 \\ E\left(\frac{\partial A}{\partial \beta}\right) &= -2\beta \\ E\left(\frac{\partial^2 A}{\partial \beta^2}\right) &= 0 \end{aligned}$$

$$\frac{\partial^2}{\partial \beta^2} \log L = -\frac{1}{2} \left[\left\{ \frac{2}{(1-\beta^2)^2} + \frac{8\beta^2}{(1-\beta^2)^3} \right\} A + \frac{4\beta}{(1-\beta^2)^2} \frac{\partial A}{\partial \beta} + \frac{1}{1-\beta^2} \frac{\partial^2 A}{\partial \beta^2} \right],$$

whence

$$E\left(\frac{\partial^2}{\partial \beta^2} \log L\right) = -\frac{1}{1-\beta^2},$$

and thus the variance of the ML estimator is given (cf. (18.60)) by

$$\text{var } \hat{\beta} = \frac{1-\beta^2}{n}. \quad (50.21)$$

For $\beta = \frac{1}{2}$ this reduces to $3/4n$ and comparison with (50.20) shows that the estimator b from (50.16) has a variance 3.6 times the optimum value.

The result is unexpected but easy to understand. The estimator of (50.16) uses only the first serial correlation and forfeits the information in the other serials which, as we have seen, all appear in the likelihood function.

Estimation in autoregressive series

50.7 For the general linear autoregressive series

$$\sum_{j=0}^k \alpha_j u_{t-j} = \varepsilon_t \quad (50.22)$$

the same kind of argument as we used in 50.4 shows that, with the usual neglect of end-effects, the ML estimators in normal variation are given by minimizing

$$\sum_{t=k+1}^n \left\{ \sum_{j=0}^k \alpha_j u_{t-j} \right\}^2$$

which gives rise to the Yule-Walker equations (47.66). Or equivalently, we can treat the estimation problem as one in ordinary regression.

The basic theorem on this subject is due to Mann and Wald (1943) who proved rigorously that asymptotically the sampling properties of least-squares estimators $\hat{\alpha}$ of α are the same as those of least-squares regression estimators in multivariate normal systems. This useful result is enough for most practical purposes. Experimental studies on series generated by rectangularly distributed ε 's, and for moderate length n of 60 terms, indicate that the Yule-Walker equations can safely be used in such cases, though it is better to correct the estimates of autocorrelations for bias by Quenouille's method—cf. 48.4.

50.8 The kind of hypothesis concerning generating schemes which we mostly wish to test concerns the comparison of an autoregressive scheme of order k with one of order $k+1$. That is to say, if we assume that the series is autoregressive, how far do we carry the regressions? From what has been said it will be evident that we can go on fitting extra terms in the regression until there is no appreciable diminution in the sum of squares. In fact, autoregressive fitting is rather simpler than ordinary regression because we do not have to face the usual problems of how to reject "insignificant" variables when the regressors are of mixed types.

Example 50.2

In Table 45.4 we gave a series of figures for the sheep population of England and Wales from 1867 to 1939. Fig. 45.4 indicates that the downward trend in these figures

Table 50.1—Residual values of the sheep series of Table 45.4 after elimination of trend by a simple nine-point moving average

Year	Residual (10,000)	Year	Residual (10,000)	Year	Residual (10,000)
1871	-176	1893	+ 34	1915	+ 19
72	-112	94	-103	16	+128
73	+ 50	95	-104	17	+ 97
74	+141	96	- 15	18	+ 69
75	+ 60	97	- 23	19	- 29
76	- 20	98	+ 17	20	-174
77	+ 12	99	+ 71	21	-107
78	+ 82	1900	+ 35	22	-142
79	+130	01	+ 16	23	-109
80	- 14	02	- 27	24	- 23
81	-166	03	- 32	25	+ 60
82	-179	04	- 49	26	+121
83	- 84	05	- 61	27	+ 94
84	+ 38	06	- 52	28	- 25
85	+ 97	07	- 24	29	- 90
86	+ 8	08	+ 68	30	- 75
87	- 5	09	+141	31	+ 72
88	-105	10	+119	32	+152
89	- 99	11	+ 66	33	+112
90	+ 35	12	- 52	34	- 64
91	+159	13	-117	35	- 87
92	+167	14	- 61		

is approximately linear. In Table 50.1 we show the residuals in this series after the elimination of trend by a simple nine-point moving average. We have to consider how far this residual series can be represented by an expression of the form

$$u_t = f(u_{t-1}, u_{t-2}, \dots) + \varepsilon_t. \quad (50.23)$$

The first ten serial correlations are as follows:

Order of correlation k	r_k	Order of correlation k	r_k
1	0.595	6	0.144
2	-0.151	7	0.203
3	-0.601	8	0.118
4	-0.537	9	0.006
5	-0.138	10	-0.078

We first of all consider what order of linear autoregressive scheme would be required. This is most easily decided in terms of partial correlations of u_t with u_{t-k} eliminating all intervening observations, and the corresponding multiple correlations determined by (27.61). We find—

lag k	Value of partial r of lag k	$\sum_{i=1}^k (1-r^2) = 1-R^2$
1	0.595	0.6460
2	-0.782	0.2509
3	0.097	0.2485
4	-0.183	0.2402
5	0.031	0.2400
6	0.014	0.2400

There is apparently no appreciable gain in representation to be obtained by taking a linear autoregression of order greater than two. Note that the high values of $|r_3|$, $|r_4|$ disappear upon partialling, whereas the small $|r_2|$ is replaced by the largest partial $|r|$.

We might, however, wish to examine the question whether curvilinear terms might improve the autoregression fit (even at the expense of rendering the model non-stationary). This is most clearly decided by drawing the scatter diagrams of u_t on u_{t-1} and of u_t on u_{t-2} , which are shown in Fig. 50.1. There is no sign, to the eye at least, of curvilinearity in this scatter of variation. We conclude that, so far as autoregressive representation is possible, it is adequate to take the Yule scheme

$$u_t = -\alpha_1 u_{t-1} - \alpha_2 u_{t-2} + \varepsilon_t, \quad (50.24)$$

in which the variance of ε is about 25 per cent (i.e. the value of $1 - R^2$ above—cf. (27.56)) of the variance of u .

The constants α_1 and α_2 are easily estimated using (47.77–8) as

$$-\alpha_1 = \frac{r_1(1-r_2)}{1-r_1^2} = 1.060 \quad -\alpha_2 = \frac{r_2-1}{1-r_1^2} + 1 = -0.782,$$

and the autoregressive equation is

$$u_t = 1.060u_{t-1} - 0.782u_{t-2}. \quad (50.25)$$

Test of fit for autoregressive schemes

50.9 It so happens that for autoregressive schemes the partial autocorrelations can be obtained directly, a fact which was used by Quenouille (1947b) to provide an ingenious test of fit.

Corresponding to (50.22) consider a variable η_t defined by

$$\sum_0^k \alpha_j u_{t+j} = \eta_t, \quad (50.26)$$

where the u 's go forward, so to speak, instead of backward in time.

We have

$$\begin{aligned} \text{cov}(\eta_t, \eta_{t+l}) &= E(\sum_i \alpha_i u_{t+i})(\sum_j \alpha_j u_{t+j+l}) \\ &= \sum_{i,j=0}^k \alpha_i \alpha_j \gamma_{|l+j-i|} \end{aligned}$$

where γ_p is the p th autocovariance of u_t ,

$$= \sum_{i=0}^k \alpha_j \sum_{j=0}^k \alpha_i \gamma_{|j+l-i|}. \quad (50.27)$$

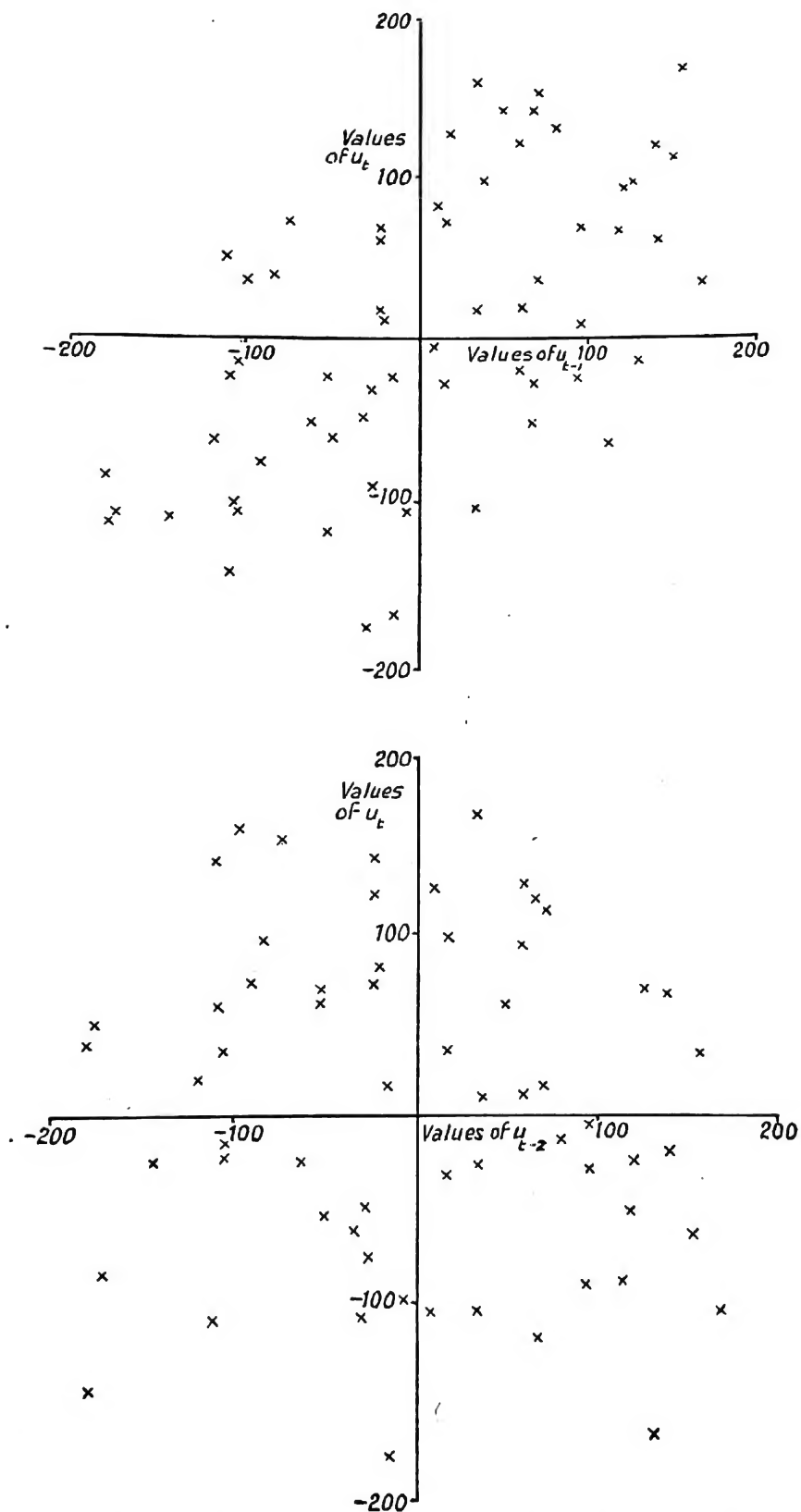


Fig. 50.1—Scatter diagrams of u_t and u_{t-1} (upper diagram) and u_t and u_{t-2} (lower diagram) for sheep data (Example 50.2)

The second summation on the right vanishes, in virtue of the Yule-Walker equations (47.66), for $l > 0$. The same result holds for $l < 0$. When $l = 0$ we find

$$\text{var } \eta_l = \text{var } \varepsilon_l. \quad (50.28)$$

Thus, if ε is a normal random variable, so is η , with the same variance. η_l depends on ε_{l+k} and previous ε 's and is therefore independent of ε_{l+k+l} for $l > 0$.

Consider now quantities q defined by

$$q_j = \frac{1}{n} \sum_{t=1}^n (\varepsilon_{t+j} \eta_t). \quad (50.29)$$

We have

$$E(q_j) = 0, \quad j > k. \quad (50.30)$$

$$\begin{aligned} \text{var } q_j &= E(\varepsilon_{t+j}^2) E(\eta_t^2) \\ &= (\text{var } \varepsilon)^2, \quad j > k. \end{aligned} \quad (50.31)$$

$$\text{cov}(q_j, q_{j+l}) = 0, \quad l \neq 0. \quad (50.32)$$

Define the quantities

$$\begin{aligned} \omega_j &= q_j / \text{var } u \\ &= \frac{1}{n \text{ var } u} \sum_{t=1}^n \varepsilon_{t+j} (\alpha_0 u_t + \dots + \alpha_k u_{t+k}), \quad j > k. \end{aligned} \quad (50.33)$$

Then each ω_j has zero mean, variance equal to $(\text{var } \varepsilon / \text{var } u)^2$, and is uncorrelated with the other ω 's.

We observe that ε_{t+j} is u_{t+j} after removal by regression of the terms $u_{t+j-1}, \dots, u_{t+j-k}$. On the other hand, η_t is u_t after the removal of u_{t+1}, \dots, u_{t+k} . Thus, for $j > k$ the correlation between ε_{t+j} and η_t , namely ω_j , is the partial correlation of terms in the series distance j apart.

For large samples we have, from (50.22) and (50.33), using sample values for the serial correlations,

$$\begin{aligned} \omega_j &= \frac{1}{n \text{ var } u} \sum_{t=1}^n (\alpha_0 u_{t+j} + \dots + \alpha_k u_{t+j-k}) (\alpha_0 u_t + \dots + \alpha_k u_{t+k}) \\ &= A_0 r_j + A_1 r_{j-1} + \dots + A_{2k} r_{j-2k}, \end{aligned} \quad (50.34)$$

where the A 's are given by

$$A_i = \sum_{j=0}^i \alpha_j \alpha_{i-j}. \quad (50.35)$$

Example 50.3

For the Yule scheme (47.74) we find, from (50.35),

$$A_0 = 1, \quad A_1 = 2\alpha_1, \quad A_2 = \alpha_1^2 + 2\alpha_2, \quad A_3 = 2\alpha_1\alpha_2, \quad A_4 = \alpha_2^2,$$

and hence, asymptotically,

$$\omega_j = r_j + 2\alpha_1 r_{j-1} + (\alpha_1^2 + 2\alpha_2) r_{j-2} + 2\alpha_1\alpha_2 r_{j-3} + \alpha_2^2 r_{j-4}, \quad j > 2, \quad (50.36)$$

is distributed with variance

$$\frac{1}{n} \left[\frac{1-\alpha_2}{1+\alpha_2} \{ (1+\alpha_2)^2 - \alpha_2^2 \} \right]. \quad (50.37)$$

In Example 50.2 we found, for the series of 65 terms of residuals in the sheep series,

$$\alpha_1 = -1.060, \quad \alpha_2 = 0.782.$$

Substitution in (50.36) then gives

$$\omega_j = r_j - 2.120r_{j-1} + 2.688r_{j-2} - 1.658r_{j-3} + 0.612r_{j-4},$$

with variance 9.69×10^{-4} .

Considered as estimators of partial correlations in series of moderate length these quantities are rather indifferent, being affected by sampling fluctuations or casual errors. They rest on the assumption, of course, that we can use sample estimators of the α 's. We find

$$\omega_3 = 0.025, \quad \omega_4 = -0.043, \quad \omega_5 = -0.001,$$

with a standard error of 0.031, and reach the same conclusion as in Example 50.2, that a second-order scheme is sufficient to account for the data.

Moving-average processes

50.10 The difficulties we remarked upon in estimation of the constants in the pure moving-average process (50.11) are obviously intensified when averages of greater extent are concerned. Asymptotic expressions for the likelihood may be derived, but the ML equations are extremely cumbrous. We shall describe a method due to Durbin (1959b) which, in effect, turns the problem into one of autoregression.

Consider, in fact, the simple model (50.11). This is equivalent to the infinite autoregression

$$u_t - \beta u_{t-1} + \beta^2 u_{t-2} - \dots = \varepsilon_t. \quad (50.38)$$

Compare this with the finite autoregressive scheme of order

$$u_t + \alpha_1 u_{t-1} + \alpha_2 u_{t-2} + \dots + \alpha_k u_{t-k} = \varepsilon_t, \quad (50.39)$$

with

$$\alpha_k = (-\beta)^k.$$

The difference lies in the remainder after $k+1$ terms of the autoregression:

$$\begin{aligned} & (-\beta)^{k+1} u_{t-k-1} + (-\beta)^{k+2} u_{t-k-2} + \dots \\ &= (-\beta)^{k+1} \{u_{t-k-1} - \beta u_{t-k-2} + \dots\} \\ &= (-\beta)^{k+1} \varepsilon_{t-k-1}. \end{aligned} \quad (50.40)$$

The variance of this term is $\beta^{2k+2} \text{var } \varepsilon$, and for $|\beta| < 1$ this tends rapidly to zero as k grows larger. Consequently the representation (50.39) can be made as close to (50.38) as we like by taking k sufficiently large (but small compared to n).

Let a_1, a_2, \dots, a_k be the least-squares estimators of $\alpha_1, \alpha_2, \dots, \alpha_k$ in (50.39). From the Mann-Wald theorem of 50.7, and (19.16), we know that the $(a - \alpha)$'s are asymptotically normal with zero mean and dispersion matrix equal to V_k^{-1}/n , where $V_k \text{var } \varepsilon$ is the dispersion matrix of the regressor variables, namely of $u_{t-1}, u_{t-2}, \dots, u_{t-k}$. This is given by the matrix of (50.12). Hence for the asymptotic distribution of a_1, \dots, a_k we have

$$dF = \frac{n^{\frac{1}{2}} |V_k|^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}k}} \exp \left[-\frac{1}{2} n \left\{ (1 + \beta^2) \sum_{j=1}^k (a_j - \alpha_j)^2 + 2\beta \sum_{j=1}^{k-1} (a_j - \alpha_j)(a_{j+1} - \alpha_{j+1}) \right\} \right] da_1 \dots da_k. \quad (50.41)$$

The expression in curly brackets, say Q , is the essential part of the likelihood function, since $|V_k|$ is $(1 - \beta^{2k+2})/(1 - \beta^2)$ (cf. Exercise 50.1). We can simplify Q to some

extent. Consider the Yule-Walker equations (47.66) in the form

$$\left. \begin{aligned} \alpha_1 \gamma_0 + \alpha_2 \gamma_1 + \dots + \alpha_k \gamma_{k-1} &= -\gamma_1 \\ \alpha_1 \gamma_1 + \alpha_2 \gamma_2 + \dots + \alpha_k \gamma_k &= -\gamma_2 \\ &\vdots \\ \alpha_1 \gamma_{k-1} + \alpha_2 \gamma_k + \dots + \alpha_k \gamma_0 &= -\gamma_k. \end{aligned} \right\} \quad (50.42)$$

Putting $\gamma_0 = (1 + \beta^2)\sigma^2$, $\sigma^2 = \text{var } \varepsilon$, $\gamma_1 = \beta\sigma^2$, we have

$$\left. \begin{aligned} (1 + \beta^2)\alpha_1 + \beta\alpha_2 &= -\beta, \\ \beta\alpha_{r-1} + (1 + \beta^2)\alpha_r + \beta\alpha_{r+1} &= 0, \quad r = 2, 3, \dots, k-1, \\ \beta\alpha_{k-1} + (1 + \beta^2)\alpha_k &= 0. \end{aligned} \right\} \quad (50.43)$$

Multiplying these equations by $-2a_j + \alpha_j$ ($j = 1, \dots, k$) in turn and adding, we find the expression

$$Q = (1 + \beta^2) \sum_{j=1}^k a_j^2 + 2\beta \sum_{j=1}^{k-1} a_j a_{j+1} + 2\beta a_1 - \beta\alpha_1. \quad (50.44)$$

Since for large k , α_1 is nearly equal to $-\beta$, this gives, on putting $a_0 = 1$,

$$Q = (1 + \beta^2) \sum_{j=1}^k a_j^2 + 2\beta \sum_{j=0}^{k-1} a_j a_{j+1} - 1. \quad (50.45)$$

The estimator of β is now given by differentiating Q and equating to zero in the usual way, which gives us

$$b = - \frac{\sum_{j=0}^{k-1} a_j a_{j+1}}{\sum_{j=1}^k a_j^2}. \quad (50.46)$$

50.11 This estimator is easily computed from the a 's, which in turn are derivable without difficulty from a regression routine. Durbin (1959b) showed, moreover, that to the first order in n (cf. Exercise 50.6)

$$\text{var } b = \frac{2}{nE(\partial^2 Q / \partial \beta^2)}. \quad (50.47)$$

In the present case

$$E\left(\frac{\partial^2 Q}{\partial \beta^2}\right) = 2 \sum_{j=0}^k a_j^2$$

which, for large k , tends to $2 \sum_{j=0}^{\infty} (-\beta)^{2j} = 2/(1 - \beta^2)$. Thus for sufficiently large k ,

$$\text{var } b = \frac{1 - \beta^2}{n}, \quad (50.48)$$

and the estimator b is asymptotically efficient.

50.12 Similar methods give acceptable results for higher-order processes, but the expressions become more complicated. We will quote without proof the main results. The process is

$$u_t = \sum_0^h \beta_j \varepsilon_{t-j} \quad (50.49)$$

with the ε 's independently and identically (but not necessarily normally) distributed. The asymptotic distribution of the least-squares estimators \mathbf{a} of α is given by

$$dF = \frac{n^{\frac{1}{2}} |B|^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}k}} \exp(-\frac{1}{2}nQ) d\mathbf{a}, \quad (50.50)$$

where $\sigma^2 \mathbf{B}$ is the dispersion matrix of u_{t-1}, \dots, u_{t-k} and

$$Q = (\mathbf{a} - \alpha)' \mathbf{B} (\mathbf{a} - \alpha) = \mathbf{a}' \mathbf{B} \mathbf{a} - 2\mathbf{a}' \mathbf{B} \alpha + \alpha' \mathbf{B} \alpha. \quad (50.51)$$

This simplifies to

$$Q = \mathbf{a}' \mathbf{B} \mathbf{a} + 2\mathbf{a}' \mathbf{c} - \alpha' \mathbf{c}, \quad (50.52)$$

where $\mathbf{c} = (c_1, \dots, c_k)$ and $c_j = \gamma_j / \sigma^2$,

$$\text{and again to} \quad Q = \mathbf{a}' \mathbf{B} \mathbf{a} + 2\mathbf{a}' \mathbf{c} + \sum_{j=1}^h \beta_j^2. \quad (50.53)$$

The estimators b of β are given by

$$\begin{bmatrix} \sum_{j=0}^k a_j^2 & \sum_{j=0}^{k-1} a_j a_{j+1} & \sum_{j=0}^{k-2} a_j a_{j+2} & \dots & \sum_{j=0}^{k-h+1} a_j a_{j+h-1} \\ \sum_{j=0}^{k-1} a_j a_{j+1} & \sum_{j=0}^k a_j^2 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{j=0}^{k-h+1} a_j a_{j+h-1} & \sum_{j=0}^{k-h+2} a_j a_{j+h-2} & \dots & \dots & \sum_{j=0}^k a_j^2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_h \end{bmatrix} = - \begin{bmatrix} \sum_{j=0}^{k-1} a_j a_{j+1} \\ \sum_{j=0}^{k-2} a_j a_{j+2} \\ \dots \\ \sum_{j=0}^{k-h} a_j a_{j+h} \end{bmatrix} \quad (50.54)$$

The asymptotic variance matrix of b is approximately

$$\frac{1}{n} \left[E \frac{\partial^2 Q}{\partial \beta_i \partial \beta_j} \right]^{-1} \quad (50.55)$$

which may be shown to be equal to U/n , where

$$U = \begin{pmatrix} 1 - \beta_h^2 & \beta_1 - \beta_{h-1} \beta_h & \beta_2 - \beta_{h-2} \beta_h & \dots & \beta_{h-1} - \beta_1 \beta_h \\ \beta_1 - \beta_{h-1} \beta_h & 1 + \beta_1^2 - \beta_{h-1}^2 - \beta_h^2 & \dots & \dots & \dots \\ \beta_2 - \beta_{h-2} \beta_h & \beta_1 + \beta_1 \beta_2 - \beta_{h-2} \beta_{h-1} - \beta_{h-1} \beta_h & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \beta_{h-1} - \beta_1 \beta_h & \dots & \dots & \dots & 1 - \beta_h^2 \end{pmatrix} \quad (50.56)$$

50.13 The foregoing results provide a basis for the construction of large-sample tests of hypotheses. For example, in (50.11), to test the hypothesis that $\beta = \beta_0$ we calculate b from (50.42) and test

$$z = \sqrt{n(b - \beta_0)(1 - \beta_0^2)^{-\frac{1}{2}}} \quad (50.57)$$

as a normal deviate $N(0, 1)$. Likewise, in the more general scheme (50.49) we test

$$v = n \sum_{i=1}^h \sum_{j=1}^h u_{ij}^{-1} (b_i - \beta_i)(b_j - \beta_j) \quad (50.58)$$

as a χ^2 variable with h d.f. Here (u_{ij}^{-1}) is the inverse of (50.56).

50.14 Again, a test of the goodness of fit of the whole model may be derived. With the simpler model (50.11) we note that nQ , where Q is given by (50.45), is distributed approximately as χ^2 with k d.fr. Substitution in (50.45) from (50.46) shows that Q can be partitioned in the form

$$Q = (1-b^2) \sum_0^k a_j^2 - 1 + (b-\beta)^2 \sum_0^k a_j^2. \quad (50.59)$$

Asymptotically $n(b-\beta)^2 \sum a_j^2$ is equivalent to the regression sum of squares in a linear regression model, the remainder being the residual sum of squares. Thus the goodness of fit of the model may be examined by testing

$$n\{(1-b^2) \sum_0^k a_j^2 - 1\} \quad (50.60)$$

as a χ^2 variable with $k-1$ d.fr.

For the more extended model (50.49) the minimum value of Q is given by

$$\sum_0^k a_j^2 + \sum_{j=1}^h b_j \sum_{i=0}^{k-j} a_i a_{i+j} - 1 \quad (50.61)$$

which can be tested in χ^2 with $k-h$ d.fr.

For details and some numerical results see Durbin (1959b). Wold (1949) had earlier suggested a more complicated test. Durbin proved for the second-order case, and conjectured a general result, to the effect that the limiting dispersion determinant of the autoregressive scheme

$\sum_0^h \alpha_j u_{t-j} = \varepsilon_t$ is the same as the limiting dispersion determinant of

the moving-average scheme $u_t = \sum_0^h \alpha_j \varepsilon_{t-j}$. This was proved by Finch (1960) and by

A. M. Walker (1961). If there is any simple explanation of this remarkable duality it remains undiscovered.

Autoregressive schemes with moving-average errors

50.15 Consider now the mixed scheme

$$\sum_{j=0}^k a_j u_{t-j} = \sum_{j=0}^h \beta_j \varepsilon_{t-j}. \quad (50.62)$$

The problem of finding efficient estimators of α 's and β 's has not been thoroughly investigated, but the most promising method, due to Durbin (1960b), seems to be to iterate in the following manner.

Suppose we have a set of values \mathbf{a} of α . We can then transform the u 's to new variables z by the autoregressive transformation

$$z_t = \sum_{j=0}^k a_j u_{t-j} \quad (50.63)$$

and estimate the constants β in the model

$$z_t = \sum \beta_j \varepsilon_{t-j}. \quad (50.64)$$

Having determined estimates \mathbf{b} of β , we can now transform

$$\sum \alpha_j u_{t-j} = \sum b_j \varepsilon_{t-j}$$

to autoregressive equations of the form

$$\sum \alpha'_j u_{t-j} = \varepsilon_t, \quad (50.65)$$

where the α' are linear functions of the α 's. We can then obtain estimates of the α 's and hence of the α' s. The cycle can then begin again until the estimates of α 's and β 's converge.

50.16 The problems of applying this procedure are twofold: to ensure that the iterative procedure converges satisfactorily, and to find a good set of starting values. The first problem does not seem to have been thoroughly examined—in time-series analysis convergence is not a property which can be assumed without extensive practical testing. As to the second, we recall that the scheme (50.62) can be closely approximated by an autoregressive scheme of large order. If we fit such a scheme, let the residuals be e_t . In (50.62) we replace ε_t by e_t and hence obtain preliminary estimates of α and β by minimizing

$$\left(\sum_{j=0}^k \alpha_j u_{t-j} - \sum_{j=1}^h \beta_j e_{t-j} \right)^2. \quad (50.66)$$

Example 50.4

Consider the model

$$u_t + \alpha u_{t-1} = \varepsilon_t + \beta \varepsilon_{t-1}. \quad (50.67)$$

Approximate by a scheme of order k giving residuals e_t

$$e_t = u_t + a_1 u_{t-1} + \dots + a_k u_{t-k}. \quad (50.68)$$

We minimize

$$\sum (u_t + \alpha u_{t-1} - \beta e_{t-1})^2$$

to obtain for estimators of α and β :

$$\sum u_t u_{t-1} + \hat{\alpha} \sum u_{t-1}^2 - \hat{\beta} \sum e_{t-1} u_{t-1} = 0, \quad (50.69)$$

$$\sum u_t e_{t-1} - \hat{\alpha} \sum u_{t-1} e_{t-1} - \hat{\beta} \sum e_{t-1}^2 = 0. \quad (50.70)$$

Substituting for e from (50.68) and replacing $\sum u_t u_{t-p}$ by $(\sum u_t^2) r_p$, we find the asymptotic expressions

$$\hat{\alpha} = - \frac{a_1 r_2 + a_2 r_3 + \dots + a_k r_{k+1}}{a_1 r_1 + a_2 r_2 + \dots + a_k r_k} \quad (50.71)$$

$$\hat{\beta} = \hat{\alpha} + \frac{r_1 + a_1 r_2 + \dots + a_k r_{k+1}}{1 + a_1 r_1 + \dots + a_k r_k}. \quad (50.72)$$

From these the iterative solution may begin.

50.17 The mixed scheme (50.62) with random values ε has the autocovariance generator (cf. 47.15 and 47.18)

$$G(z) = \frac{(\sum \beta_j z^j)(\sum \beta_j z^{-j})}{(\sum \alpha_j z^j)(\sum \alpha_j z^{-j})} \quad (50.73)$$

and the corresponding spectral density is given (cf. 47.14) by

$$w(\alpha) = \frac{(\sum \beta_j e^{i\alpha j})(\sum \beta_j e^{-i\alpha j})}{(\sum \alpha_j e^{i\alpha j})(\sum \alpha_j e^{-i\alpha j})}. \quad (50.74)$$

From estimates of the α 's and β 's we can then determine the estimated spectral density. It is more relevant, perhaps, to consider whether the α 's and β 's can be estimated from the observed spectrum. The question has been examined by Durbin (1961).

50.18 A more general assault on the problems of hypothesis testing and estimation has been made by Whittle in a series of papers, particularly 1951, 1953a and 1953b. The methods, based for the most part on Maximum Likelihood considerations, are penetrating and the results of considerable generality for stationary time-series with continuous spectra. They are, however, not very suitable for numerical work.

Some multivariate extensions

50.19 Suppose that we have p series, u_i , $i = 1, 2, \dots, p$, observed at n intervals of time or, in the continuous case, defined over a certain time-period. The value of u_i at time t will be denoted by u_{it} , and the set of $p \times n$ values of u by a matrix \mathbf{u} . As in the univariate case, we regard this as the realization of a process, and our basic object is to determine what kind of a process it is and what are its parameters.

In general, any row vector of \mathbf{u} , considered as a single series, may contain trend, seasonal, or oscillatory movements. However, it makes for almost unmanageable complication to try to dissect each vector *simultaneously* into its constituents. We shall assume that trend and seasonal movements have been removed, leaving us with a multivariate stationary complex \mathbf{u} . We follow Quenouille (1957).

50.20 The covariance of u_{it} and u_{jt-s} will be written $\gamma_{(ij)s}$ and the corresponding correlation by $\rho_{(ij)s}$. The analogous observed quantities are $c_{(ij)s}$, $r_{(ij)s}$. For any given s there are $\frac{1}{2}p(p+1)$ of these quantities arrayable in a square matrix which we write γ_s , ρ_s , c_s or r_s as the case may be. In the univariate case $\rho_s = \rho_{-s}$, but clearly $E(u_{it}u_{jt+s})$ is not equal to $E(u_{it}u_{jt-s})$ but to $E(u_{i,t-s}u_{jt})$. Hence we have

$$\gamma_s = \gamma_s'. \quad (50.75)$$

As usual with multivariate extensions, the number of parameters and estimators increase rapidly with p . We shall refer to $\gamma_{(ij)s}$ as the *cross-covariance* of u_i and u_j for lag s . Likewise $\rho_{(ij)s}$ is a *cross-correlation*. Where necessary to distinguish between sample and parental values we shall use those words, although they may often be omitted when the symbols themselves make it clear which is under reference.

50.21 As in the univariate case, we shall be concerned with three types of model, autoregressive, moving-average, and mixed autoregressive-moving-average systems. Let ε_{it} be a series of independent random elements. Corresponding to the univariate case

$$u_t = \sum_{s=0}^l \beta_s \varepsilon_{t-s}$$

we now have

$$u_{it} = \sum_{s=0}^l \sum_{j=1}^p \beta_{ijs} \varepsilon_{j,t-s}, \quad (50.76)$$

and a corresponding autoregressive scheme

$$\sum_{s=0}^k \sum_{j=1}^p \alpha_{ijs} u_{j,t-s} = \varepsilon_{it}. \quad (50.77)$$

Writing D for a shift operator such that $Du_t = u_{t-1}$, we may express (50.76) in the form

$$u_{it} = \sum_{s=0}^l \sum_{j=1}^p \beta_{ijs} D^s \varepsilon_{jt}, \quad (50.78)$$

or, in matrix form,

$$\mathbf{u}_t = \sum_{s=0}^l \mathbf{B}_s D^s \boldsymbol{\epsilon}_t \quad (50.79)$$

$$= \mathbf{B}(D) \boldsymbol{\epsilon}_t, \quad (50.80)$$

where

$$\mathbf{B}(D) = \sum_{s=0}^l \mathbf{B}_s D^s. \quad (50.81)$$

Likewise, for the autoregressive scheme of (50.77) we may write

$$\boldsymbol{\epsilon}_t = \sum_{s=0}^k \mathbf{A}_s D^s \mathbf{u}_t = \mathbf{A}(D) \boldsymbol{\epsilon}_t, \quad (50.82)$$

where

$$\mathbf{A}(D) = \sum_{s=0}^k \mathbf{A}_s D^s. \quad (50.83)$$

We may write the solution of (50.82)

$$\mathbf{u}_t = \mathbf{A}^{-1}(D) \boldsymbol{\epsilon}_t.$$

The terms in \mathbf{A} and \mathbf{B} are polynomials in D . We also define

$$\boldsymbol{\gamma} = \sum_{s=-\infty}^{\infty} \boldsymbol{\gamma}_s D^s. \quad (50.84)$$

50.22 Without loss of generality, we will choose the scales so that the ϵ_i have zero mean, the same variance for all i , say σ^2 , and are uncorrelated. Then we have

$$\boldsymbol{\gamma}_s = E(\mathbf{u}_t \mathbf{u}_{t-s}').$$

Substituting from (51.79), we find

$$\begin{aligned} \boldsymbol{\gamma}_s &= E\left(\sum_{j=0}^l \mathbf{B}_j D^j \boldsymbol{\epsilon}_t\right)\left(\sum_{m=0}^l \mathbf{B}_m D^m \boldsymbol{\epsilon}_{t-s}\right)' \\ &= \sum_{j=0}^l \sum_{m=0}^l \mathbf{B}_j \mathbf{B}_m' E(\boldsymbol{\epsilon}_{t-j} \boldsymbol{\epsilon}_{t-s-m}') \\ &= r^2 \sum_{s,m} \mathbf{B}_{s+m} \mathbf{B}_m' \\ &= \text{coeff. of } D^s \text{ in } \sigma^2 (\sum \mathbf{B}_j D^j) (\sum \mathbf{B}_j' D^{-j}). \end{aligned} \quad (50.85)$$

Hence we may write

$$\boldsymbol{\gamma} = \mathbf{B}(D) \mathbf{B}'(D^{-1}). \quad (50.86)$$

Likewise, for the autoregressive equation

$$\boldsymbol{\gamma} = \mathbf{A}^{-1}(D) \mathbf{A}'^{-1}(D^{-1}). \quad (50.87)$$

It is easy to show that for a mixed scheme

$$\begin{aligned} \mathbf{A} \mathbf{u}_t &= \mathbf{B} \boldsymbol{\epsilon}_t \\ \boldsymbol{\gamma} &= \mathbf{A}^{-1}(D) \mathbf{B}(D) \mathbf{B}'(D^{-1}) \mathbf{A}'^{-1}(D^{-1}). \end{aligned} \quad (50.88)$$

These are the multivariate analogues of (50.73). In them the D 's may be regarded as dummy variables equivalent to what we have formerly written as z . Equations (50.86)–(50.88) are, in fact, covariance generating functions.

50.23 For the autoregressive scheme we have natural generalizations of the Yule–

Walker equations (47.66). Postmultiplying (50.82) by \mathbf{u}'_{t+q} and taking expectations,

$$\begin{aligned} E \sum_{s=0}^k \mathbf{A}_s D^s \mathbf{u}_t \mathbf{u}'_{t+q} &= \sum \mathbf{A}_s E(\mathbf{u}_{t-s} \mathbf{u}'_{t+q}) \\ &= \sum_{s=0}^k \mathbf{A}_s \gamma_{q+s} = 0, \quad q > 0. \end{aligned} \quad (50.89)$$

The solution of these equations is not, however, an easy matter.

Degeneracy

50.24 Apart from the ordinary problems which we encountered in the univariate case, there are two further complications for multivariate series.

In the first place, there may exist linear relations among the variables, in which case the matrices \mathbf{A} or \mathbf{B} may become degenerate. Steps must then be taken to remove some of the variables

Example 50.5 (Quenouille, 1957)

Consider

$$\left. \begin{aligned} u_{1t} &= \varepsilon_{1t} + \varepsilon_{2, t-1} \\ u_{2t} &= \varepsilon_{1t} + \varepsilon_{2t} \\ u_{3t} &= \varepsilon_{2t} + \varepsilon_{2, t-1} \end{aligned} \right\}. \quad (50.90)$$

We have

$$\mathbf{B} = \begin{pmatrix} 1 & D \\ 1 & 1 \\ 0 & 1+D \end{pmatrix}, \quad (50.91)$$

and from (50.86)

$$\begin{aligned} \gamma &= \mathbf{B}(D) \mathbf{B}'(D^{-1}) \\ &= \begin{pmatrix} 2 & 1+D & 1+D \\ 1+D^{-1} & 2 & 1+D^{-1} \\ 1+D^{-1} & 1+D & 2+D+D^{-1} \end{pmatrix}. \end{aligned} \quad (50.92)$$

It is then found that $|\gamma| = 0$ and the matrix has rank 2. There must then be a linear relation among the variables. In this case we can almost determine it by inspection, but formally we should look for the zero latent root of γ and its associated latent vector. The latter is proportional to $1+D$, $-(1+D)$, $1-D$ and the relation is therefore

or

$$(1+D)u_{1t} - (1+D)u_{2t} + (1-D)u_{3t} = 0$$

$$u_{1t} - u_{2t} + u_{3t} = -u_{1, t-1} + u_{2, t-1} + u_{3, t-1}. \quad (50.93)$$

50.25 Degeneracies are, in practice, the exception rather than the rule, and when they occur can be dealt with fairly easily in the manner of Example 50.5. More important, and more difficult to deal with, is the fact that equation (50.88) does not determine \mathbf{A} and \mathbf{B} uniquely, however good our estimators of γ .

Write temporarily \mathbf{F} for $\mathbf{A}^{-1}\mathbf{B}$. Then (50.88) is equivalent to

$$\gamma = \mathbf{F}(D) \mathbf{F}'(D^{-1}). \quad (50.94)$$

If ϕ is any diagonal matrix for which the i th diagonal element is $\phi_i(D)/\phi_i(D^{-1})$, if ψ is any diagonal matrix with diagonal elements $\psi_i(D)/\psi_i(D^{-1})$, and if J is any matrix such that $JJ' = I$, it is easily seen by substitution in (50.94) that $F(D)\phi(D)J\psi(D)$ may replace $F(D)$. Thus, many different schemes may give rise to the same covariance matrix.

Example 50.6 (Quenouille, 1957)

Consider the matrices

$$F_1 = \begin{pmatrix} 2+D & D \\ 1 & 6+D \end{pmatrix}, \quad F_2 = \begin{pmatrix} 1+2D & -D \\ 5 & 3+2D \end{pmatrix}$$

$$F_3 = \frac{1}{5} \begin{pmatrix} 2+7D & 4+9D \\ -11-6D & 28+3D \end{pmatrix}, \quad F_4 = \frac{1}{17} \begin{pmatrix} 14+37D & -5-12D \\ 55+30D & 1+84D \end{pmatrix}.$$

It can easily be verified that

$$|F_1| = (4+D)(3+D), \quad |F_2| = (3+D)(1+4D)$$

$$|F_3| = (4+D)(1+3D), \quad |F_4| = (1+4D)(1+3D),$$

and that for each F

$$\gamma = \begin{pmatrix} 6+2D+2D^{-1} & 3+7D \\ 3+7D^{-1} & 38+6D+6D^{-1} \end{pmatrix}.$$

Furthermore, if we postmultiply the F 's respectively by orthogonal matrices J_1, J_2, J_3, J_4 , γ is unaltered.

50.26 It remains for consideration whether all the possible solutions of (50.84) are acceptable; for example, whether they all provide stationary series. So far as is known, some fairly stringent conditions must be imposed before we can derive a unique solution. The following treatment is due to Phillips (1959).

We consider the mixed scheme with independent residuals and assume (1) that A is non-degenerate (which we can always ensure as in Example 50.5) and (2) that $|A|$, a polynomial in D , has *different* roots $\lambda_1, \lambda_2, \dots, \lambda_m$. Then if $\alpha(D)$ is the adjoint of $A(D)$, we may write (50.88) as

$$\gamma = \frac{\alpha(D)B(D)B'(D^{-1})\alpha'(D^{-1})}{|A(D)| |A'(D^{-1})|}. \quad (50.95)$$

Expressing $|A|$ as the product $\prod_{r=1}^m (D - \lambda_r)$, we see that the right-hand side may be expressed in partial fractions

$$\sum_{r=1}^m \frac{K_r}{D - \lambda_r} + \sum_{r=1}^m \frac{K'_r}{D^{-1} - \lambda_r}, \quad (50.96)$$

where K_r is a $p \times p$ matrix given, according to the usual theory of partial fractions, by

$$K_r = \left[\frac{(D - \lambda_r)\alpha(D)B(D)B'(D^{-1})\alpha'(D^{-1})}{|A(D)| |A'(D^{-1})|} \right]_{D=\lambda_r}. \quad (50.97)$$

In (50.96) we do not want terms in positive powers of D , which implies the condition that $|B|$ is of lower degree than $|A|$.

For a simple root λ_r the matrix $A(\lambda_r)$ is simply degenerate, and its adjoint $\alpha(\lambda_r)$ is of unit rank—a known result in matrix theory. We may then write

$$\alpha(\lambda_r) = k_r k_r', \quad (50.98)$$

where \mathbf{k}_r is a $(p \times 1)$ column vector and \mathbf{k}_r is a $(1 \times p)$ row vector satisfying

$$\mathbf{A}(\lambda_r)\mathbf{k}_r = 0 \quad (50.99)$$

$$\mathbf{k}_r\mathbf{A}(\lambda_r) = 0. \quad (50.100)$$

In point of fact \mathbf{K}_r itself is of unit rank. For if we define a $(1 \times p)$ row vector \mathbf{l}_r by

$$\mathbf{l}_r = \frac{(D - \lambda_r)\mathbf{k}_r\mathbf{B}(D)\mathbf{B}'(D^{-1})\boldsymbol{\alpha}'(D^{-1})}{|A(D)| |A'(D^{-1})|} \quad (50.101)$$

we find on substituting (50.98) in (50.97) that

$$\mathbf{K}_r = \mathbf{k}_r\mathbf{l}_r. \quad (50.102)$$

Now from (50.99) we have

$$\mathbf{A}(\lambda_r)\mathbf{K}_r = 0, \quad r = 1, 2, \dots, m. \quad (50.103)$$

Given, then, the covariance matrix γ , we express it in partial fractions and hence determine \mathbf{K}_r and λ_r . We can thus derive the set of equations (50.103). The question is whether this set is enough to determine the coefficients in \mathbf{A} uniquely.

50.27 Consider first of all the case when all the scalar equations in $\mathbf{A}\mathbf{u}_t = \mathbf{B}\mathbf{e}_t$ are of the same order v and cannot be reduced to lower order. We now impose two further conditions, (a) that the elements in the leading diagonal of \mathbf{A} are of degree v but that non-diagonal elements are of degree $v-1$ at most (this means, among other things, that $|A|$ is not zero); (b) that the elements of the corresponding row in \mathbf{B} are of lower degree than v (this means that no terms in λ arise as numerators in (50.96)).

Without loss of generality we may suppose that the coefficient of D^v in each diagonal term in \mathbf{A} is unity. $|A|$ is of degree pv which is therefore equal to m . Any given row in \mathbf{A} then has pv coefficients to be determined, and equation (50.103), for $m = pv$ values of r , provides a set of non-homogeneous independent equations. Thus the coefficients are uniquely determined.

When \mathbf{A} is determined we find $\mathbf{B}(D)\mathbf{B}'(D^{-1})$ from

$$\mathbf{B}(D)\mathbf{B}'(D^{-1}) = \sum_{r=1}^m \frac{\mathbf{A}(D)\mathbf{K}_r\mathbf{A}'(D^{-1})}{D - \lambda_r} + \sum_{r=1}^m \frac{\mathbf{A}(D)\mathbf{K}_r'\mathbf{A}'(D^{-1})}{D^{-1} - \lambda_r}, \quad (50.104)$$

which is derived from (50.95) and (50.96). There remains an indeterminacy for \mathbf{B} itself. This can be resolved only by extraneous information.

50.28 If the equations in the system are not all of the same order we require still one further assumption to identify \mathbf{A} . Let the equations in $\mathbf{A}\mathbf{u}_t = \mathbf{B}\mathbf{e}_t$ be arranged such that the first equation is of lowest order and any subsequent equation is not of lower order than its predecessor. Then if \mathbf{A} and \mathbf{B} satisfy (50.88), so do $\boldsymbol{\mu}\mathbf{A}$ and $\boldsymbol{\mu}\mathbf{B}$, where $\boldsymbol{\mu}$ is an arbitrary matrix of constants. We can add any row of \mathbf{A} to a later row without violating the condition that the non-diagonal elements be of lower degree than the diagonal elements. But we cannot add to a preceding row. Thus $\boldsymbol{\mu}$ can only be a triangular matrix with zeros above the diagonal.

We can make the system identifiable if we are prepared to assume that the elements ε are not correlated from one equation to another, that is to say that \mathbf{B} is diagonal. For then $\mathbf{B}(D)\mathbf{B}'(D^{-1})$ is diagonal, and hence the non-diagonal elements of $\boldsymbol{\mu}\mathbf{B}\mathbf{B}'\boldsymbol{\mu}'$

are zero. Writing b_{ii} for the diagonal elements of $\mathbf{B}\mathbf{B}'$, the non-diagonal elements of $\mu\mathbf{B}(D)\mathbf{B}'(D^{-1})\mu'$ above the diagonal are found to be

$$\begin{array}{ccccccc} \mu_{11}b_{11}\mu_{21} & \mu_{11}b_{11}\mu_{31} & \mu_{11}b_{11}\mu_{41} & \cdot & \cdot & \cdot & \cdot \\ & \mu_{21}b_{11}\mu_{31} + \mu_{22}b_{22}\mu_{32} & \mu_{21}b_{11}\mu_{41} + \mu_{22}b_{22}\mu_{42} & \cdot & \cdot & \cdot & \cdot \\ & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}$$

Since μ_{11} cannot vanish, μ_{21} , μ_{31} , etc. must do so; and hence μ_{32} , μ_{42} , etc.; and so on. Hence μ is diagonal and the equations are identifiable.

50.29 It will be evident enough that the problems associated with identifiability are formidable. We can, at least on a heuristic basis, estimate the covariance matrix γ and hence the product $\mathbf{A}^{-1}(D)\mathbf{B}(D)\mathbf{B}'(D^{-1})\mathbf{A}'^{-1}(D^{-1})$. To proceed thence to the individual coefficients in \mathbf{A} and \mathbf{B} requires conditions on the problem which are not always easy to verify; and in any case appeal to extraneous knowledge is sometimes necessary to reach determinacy. One of the major outstanding problems of multivariate temporal systems, in fact, is to ensure that a model is unique; and this apart from sampling considerations.

Cross-spectra

50.30 Just as we may consider the cross-correlations of series and obtain what might be called cross-correlograms, so we may examine the extension of spectrum analysis to the simultaneous variation of series.

For any pair of series, say u_1 and u_2 , we have a set of cross-correlations $\rho_{(12)s}$, $s = -\infty, \dots, \infty$ and, in extension of the spectrum of a single series (47.21), may define a spectral density

$$w_{12}(\alpha) = \sum_{-\infty}^{\infty} \rho_{(12)s} \exp(is\alpha). \quad (50.105)$$

There is a corresponding spectral function $W(\alpha)$ defined over the range 0 to π . Conversely, as at (47.23),

$$\rho_{(12)s} = (1/\pi) \int_0^\pi W(\alpha) \exp(-is\alpha) d\alpha. \quad (50.106)$$

In univariate formulae, owing to the symmetry typified by $\rho_s = \rho_{-s}$, sine terms disappear from expressions relating spectral density to covariances or correlations. In the multivariate case, $\rho_{(12)s}$ is not the same as $\rho_{(12)(-s)}$. Expansion of (50.105) gives us

$$\begin{aligned} w_{12}(\alpha) &= 1 + \sum_1^{\infty} \{\rho_{(12)s} \cos s\alpha + \rho_{(12)(-s)} \cos s\alpha\} + i \sum_1^{\infty} \{\rho_{(12)s} \sin s\alpha - \rho_{(12)(-s)} \sin s\alpha\} \\ &= c(\alpha) + iq(\alpha), \quad \text{say.} \end{aligned} \quad (50.107)$$

50.31 The quantity $c(\alpha)$ is called the *co-spectrum* or *co-spectral density*. $q(\alpha)$ is called the *quadrature spectrum* or *quadrature spectral density*. Sometimes both these quantities are plotted against α . The sum of squares $c^2 + q^2$ is called the *amplitude* of the spectrum. The standardized quantity

$$C(\alpha) = \frac{c^2(\alpha) + q^2(\alpha)}{w_1(\alpha)w_2(\alpha)}, \quad (50.108)$$

where w_1 and w_2 are the spectral densities of u_1 and u_2 , is called the *coherence*.

Phase relationships in the series are studied by three types of diagram: the *phase diagram*, plotting $\psi(\alpha)$ against α , where

$$\psi(\alpha) = \arctan \left\{ \frac{q(\alpha)}{c(\alpha)} \right\}, \quad (50.109)$$

the *Argand diagram*, which plots $c(\alpha)/w_1(\alpha)$ as abscissa against $q(\alpha)/w_1(\alpha)$ as ordinate; and the *gain diagram*, plotting $R_{12}^2(\alpha)$ against α , where

$$R_{12}^2(\alpha) = \frac{w_1(\alpha)C(\alpha)}{w_2(\alpha)}. \quad (50.110)$$

A good computer programme will calculate and graph the quantities required. (50.108) and (50.110) are analogues of correlation and regression coefficients. Some further details are given by Granger and Hatanaka (1964).

50.32 For multivariate series of the autoregressive or moving-average types there is a straightforward generalization of the relation between the covariance-generating function and the spectral density. We have, in fact,

$$w(\alpha) = \mathbf{A}^{-1}(e^{i\alpha}) \mathbf{B}(e^{i\alpha}) \mathbf{B}'(e^{-i\alpha}) \mathbf{A}'^{-1}(e^{-i\alpha}), \quad (50.111)$$

but this is not, in practice, a very useful formula.

The generalization of spectra and cross-spectra to *polyspectra* for k -dimensional time-series is discussed by Brillinger (1965).

Example 50.7

To give some idea of what cross-correlations and cross-spectra look like we take an artificial series constructed by Quenouille (1957), reproduced in Table 50.2. The series was constructed from

$$\left. \begin{aligned} u_{1t} &= u_{1,t-1} - 0.1u_{2,t-1} + \varepsilon_{1t} \\ u_{2t} &= 0.2u_{1,t-1} + u_{2,t-1} - 0.3u_{3,t-1} + \varepsilon_{2t} \\ u_{3t} &= 0.9u_{2,t-1} + \varepsilon_{3t} \end{aligned} \right\} \quad (50.112)$$

The ε 's are rectangular random variables ranging from -49 to $+49$.

Table 50.3 gives, for $s = 0$ to 5 , the theoretical covariances γ_s and the observed covariances c_s . The serial correlations up to order 25 are given in Table 50.4.

In Figs. 50.2 and 50.3 we have graphed the logarithms of the spectral ordinates of the three series and the logarithms of the amplitudes of the cross-spectra. The series are effectively Markovian. Their cross-spectra exhibit much the same pattern as the schemes themselves.

50.33 It hardly needs to be stated that problems of estimation and hypothesis testing for multivariate series are much more complicated than in the univariate case, which themselves, as we have seen, are far from simple.

Scrutiny of the correlogram or power spectrum for individual series will usually suggest whether the series is stationary, whether a Markoff scheme is likely to be sufficient, or whether some more elaborate scheme may be required to explain observations. The basic elements used in deciding such questions are the serial correlations

Table 50.2—Artificial series (see text)

t	u_{1t}	u_{2t}	u_{3t}	t	u_{1t}	u_{2t}	u_{3t}	t	u_{1t}	u_{2t}	u_{3t}	t	u_{1t}	u_{2t}	u_{3t}
1	118	155	150	26	-18	11	-18	51	199	162	84	76	82	-16	-11
2	102	128	114	27	7	-15	3	52	137	139	142	77	-74	-34	1
3	131	148	122	28	20	28	-58	53	126	100	142	78	-117	-22	13
4	129	99	132	29	-29	6	27	54	73	130	70	79	-97	-10	-13
5	72	97	114	30	-20	39	53	55	76	124	121	80	-51	-64	4
6	78	63	123	31	11	-12	68	56	46	152	124	81	-42	-101	-34
7	117	33	91	32	57	7	-20	57	31	173	135	82	-29	-90	-131
8	134	44	67	33	65	-3	-37	58	20	165	160	83	-1	-88	-105
9	174	73	12	34	98	9	43	59	-54	109	162	84	-28	-29	-71
10	146	63	104	35	102	19	35	60	-29	60	52	85	-1	-46	-48
11	94	54	38	36	139	61	-15	61	-32	68	64	86	26	17	-39
12	62	27	42	37	128	58	27	62	-82	72	95	87	62	70	-29
13	56	2	17	38	101	64	85	63	-110	3	55	88	81	107	33
14	39	37	2	39	132	91	85	64	-76	34	-23	89	77	149	103
15	37	0	70	40	95	89	95	65	-69	-83	15	90	57	172	126
16	8	42	45	41	43	94	73	66	-92	-51	-35	91	52	156	169
17	-25	-79	-54	42	43	41	91	67	-90	-100	-46	92	54	72	172
18	17	-104	-103	43	62	-6	0	68	-71	-122	-166	93	82	11	19
19	38	-65	-109	44	108	-11	16	69	-45	-137	-88	94	45	45	2
20	42	2	-18	45	88	-19	2	70	-74	-125	-132	95	40	53	54
21	49	6	-10	46	134	36	-2	71	-30	-92	-73	96	33	11	21
22	52	21	38	47	139	51	-8	72	-46	-70	-43	97	46	46	-28
23	51	-20	29	48	153	69	52	73	-48	-93	-72	98	-	108	48
24	11	-16	19	49	178	108	82	74	-88	-66	-78	99	4	94	94
25	2	-29	-26	50	181	127	129	75	-70	-48	-15	100	37	61	48

THE ADVANCED THEORY OF STATISTICS

Table 50.3—Values of γ_s and c_s for the series of Table 50.2

(Values underlined are the values of the determinants of the corresponding matrices)

s	γ_s	c_s
0	$\begin{bmatrix} 7,878.14 & 4,392.92 & 3,396.37 \\ 4,392.92 & 6,191.72 & 4,010.45 \\ 3,396.37 & 5,010.45 & 5,831.96 \end{bmatrix}$ $\underline{5.2245 \times 10^{10}}$	$\begin{bmatrix} 6,896.05 & 4,413.37 & 3,511.74 \\ 4,413.37 & 6,625.07 & 5,573.84 \\ 3,511.74 & 5,573.84 & 6,161.75 \end{bmatrix}$ $\underline{3.8320 \times 10^{10}}$
1	$\begin{bmatrix} 7,438.85 & 3,773.75 & 2,895.33 \\ 4,949.64 & 5,567.17 & 3,940.14 \\ 3,935.63 & 5,572.55 & 4,509.41 \end{bmatrix}$ $\underline{1.4106 \times 10^{10}}$	$\begin{bmatrix} 6,487.04 & 3,877.69 & 3,097.75 \\ 4,822.82 & 5,984.09 & 4,624.31 \\ 3,982.18 & 5,944.23 & 4,979.27 \end{bmatrix}$ $\underline{8.2402 \times 10^9}$
2	$\begin{bmatrix} 6,943.89 & 3,217.03 & 2,501.31 \\ 5,251.32 & 4,650.15 & 3,166.38 \\ 4,454.67 & 5,010.45 & 3,546.13 \end{bmatrix}$ $\underline{3.8087 \times 10^9}$	$\begin{bmatrix} 6,053.67 & 3,384.15 & 2,784.75 \\ 5,008.42 & 5,175.63 & 3,681.88 \\ 4,421.45 & 5,315.40 & 3,925.26 \end{bmatrix}$ $\underline{3.4807 \times 10^9}$
3	$\begin{bmatrix} 6,418.76 & 2,752.01 & 2,184.68 \\ 5,303.69 & 3,790.42 & 2,602.81 \\ 4,726.19 & 4,185.14 & 2,849.74 \end{bmatrix}$ $\underline{1.0283 \times 10^9}$	$\begin{bmatrix} 5,589.50 & 2,865.88 & 2,292.22 \\ 4,877.75 & 4,224.71 & 2,932.37 \\ 4,510.29 & 4,549.42 & 3,332.35 \end{bmatrix}$ $\underline{2.6326 \times 10^9}$
4	$\begin{bmatrix} 5,888.39 & 2,372.97 & 1,924.39 \\ 5,169.59 & 3,085.29 & 2,184.82 \\ 4,773.33 & 3,411.38 & 2,342.53 \end{bmatrix}$ $\underline{2.7765 \times 10^8}$	$\begin{bmatrix} 5,128.65 & 2,388.92 & 1,928.00 \\ 4,698.35 & 3,511.09 & 2,359.65 \\ 4,376.84 & 3,819.96 & 2,824.63 \end{bmatrix}$ $\underline{2.5781 \times 10^9}$
5	$\begin{bmatrix} 5,371.43 & 2,064.44 & 1,705.91 \\ 4,915.27 & 2,536.47 & 1,866.94 \\ 4,652.63 & 2,776.76 & 1,966.34 \end{bmatrix}$ $\underline{7.4966 \times 10^7}$	$\begin{bmatrix} 4,602.27 & 2,096.25 & 1,825.86 \\ 4,549.36 & 2,770.88 & 1,835.51 \\ 4,211.30 & 3,140.85 & 2,009.68 \end{bmatrix}$ $\underline{9.1749 \times 10^8}$

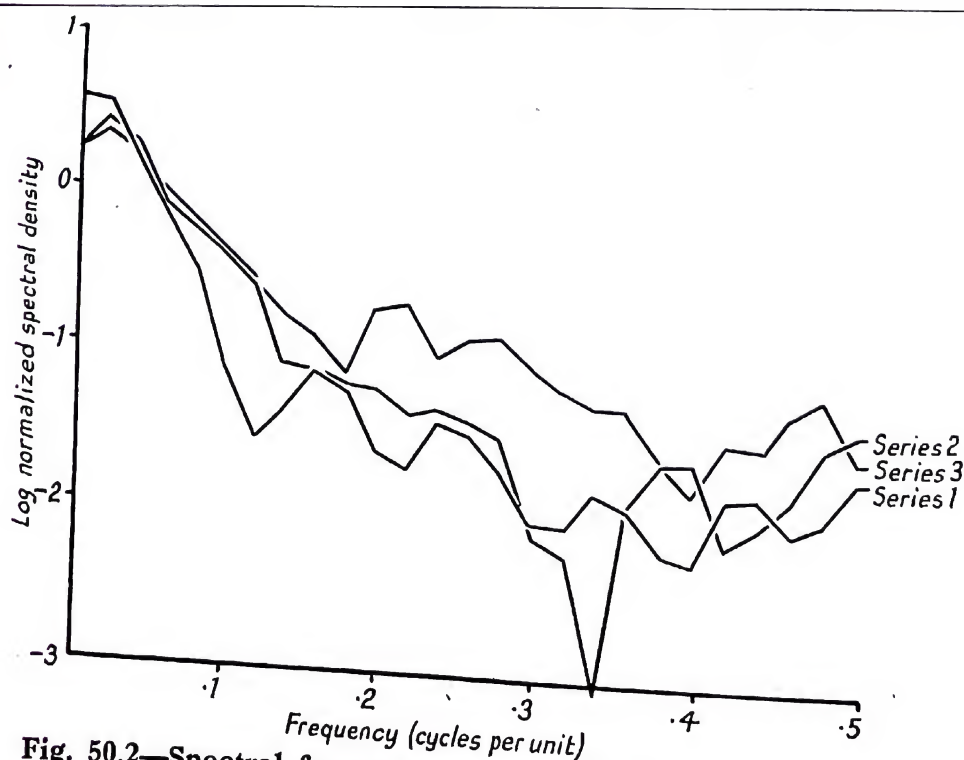


Fig. 50.2—Spectral functions of the three series of Table 50.2
The ordinate is the logarithm to base 10 of the spectral density divided by the variance of the series.

Table 50.4—Serial and cross-correlations of the series of Table 50.2

Order of correlation	Correlations (decimal points omitted)								
	Series 1 auto	Series 2 auto	Series 3 auto	2 leading 1	3 leading 1	1 leading 2	3 leading 2	1 leading 3	2 leading 3
0	1000	1000	1000	598	459	598	855	459	855
1	933	898	787	677	551	512	929	389	693
2	860	769	600	714	635	434	825	341	534
3	782	617	496	697	655	354	697	258	408
4	707	500	404	671	638	278	574	194	307
5	617	377	257	650	614	229	457	177	218
6	525	276	169	603	612	190	309	127	161
7	436	201	113	544	557	146	213	080	113
8	353	159	080	445	476	102	152	072	059
9	281	093	054	351	392	072	126	031	023
10	227	053	016	287	304	033	066	-017	-025
11	201	-014	-026	233	266	-012	034	-058	-116
12	174	-090	-099	155	219	-070	001	-085	-179
13	117	-177	-116	086	144	-116	-046	-143	-232
14	068	-258	-110	015	083	-172	-091	-216	-283
15	005	-316	-162	-036	031	-252	-159	-271	-350
16	-045	-361	-260	-096	-006	-318	-218	-305	-374
17	-110	-404	-276	-120	-059	-362	-253	-357	-388
18	-179	-419	-271	-137	-063	-416	-270	-425	-397
19	-259	-438	-234	-192	-049	-471	-276	-474	-412
20	-323	-462	-330	-279	-094	-516	-291	-493	-429
21	-381	-459	-382	-345	-159	-535	-358	-501	-482
22	-434	-535	-414	-391	-214	-545	-411	-497	-493
23	-470	-548	-401	-431	-259	-538	-448	-507	-485
24	-467	-531	-406	-485	-321	-517	-459	-465	-444
25	-473	-476	-397	-501	-345	-476	-449	-457	-407

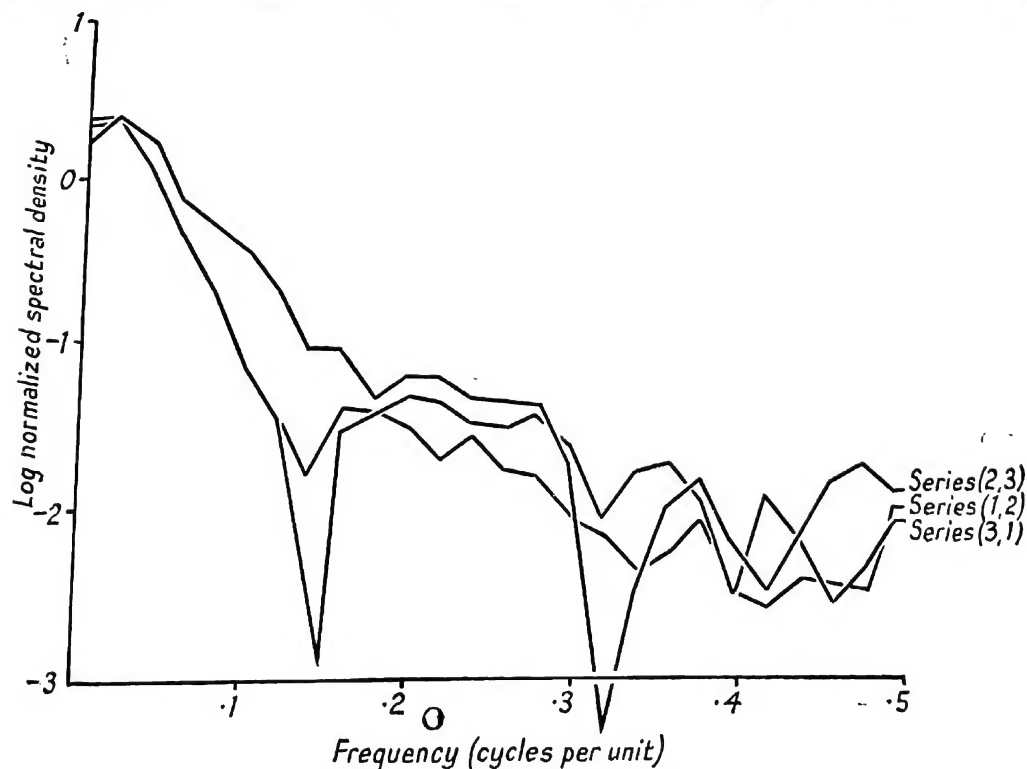


Fig. 50.3—Amplitude of cross-spectra of the three series of Table 50.2

The ordinate is the logarithm to base 10 of the cross-spectral density divided by the square root of the product of the variances of the corresponding series.

or serial covariances. We may expect that, just as in multivariate theory the dispersion determinant takes over the role of the variance, so in multivariate time-series adequate tests will be based on autocovariance or autocorrelation determinants. The exact sampling theory of such quantities has not been developed and we must be content, in the present state of knowledge, with somewhat imprecise, though intuitively reasonable, procedures.

50.34 Let us begin, then, with a consideration of the covariance determinants $|\gamma_s|$. If the scheme is one of moving averages, all such determinants vanish from some value of s , say l , onwards. If it is autoregressive, there are relations between the successive matrices γ . For example, if the scheme is of Markoff type,

$$|\gamma_s| = \beta^s |\gamma_0| \quad (50.113)$$

(cf. (47.72)) where

$$\beta = -|A_1|/|A_0|, \quad (50.114)$$

and if it is of the Yule type (cf. Example 47.8),

$$\begin{vmatrix} \gamma_{s+1} & \gamma_s \\ \gamma_s & \gamma_{s-1} \end{vmatrix} = \beta^s \begin{vmatrix} \gamma_1 & \gamma_0 \\ \gamma_0 & \gamma_1' \end{vmatrix} \quad (50.115)$$

where

$$\beta = -|A_2|/|A_0|. \quad (50.116)$$

Unfortunately these relations do not work well in practice because of the high degree of sampling variation which obscures the true facts.

For example, with the series of Example 50.7 the values of $|c_s|$, and those of $|\gamma_s|$ for $s = 0, \dots, 5$, are

s	$ \gamma_s $	$ c_s $
0	5.225×10^{10}	3.832×10^{10}
1	1.411×10^{10}	8.240×10^9
2	3.809×10^9	3.481×10^9
3	1.028×10^9	2.633×10^9
4	2.777×10^8	2.578×10^9
5	7.497×10^7	9.175×10^8

The values fluctuate too much to provide a very clear guide.

50.35 A further possibility is to consider the ratios $\gamma_s \gamma_{s-1}^{-1}$. For instance, with a Markoff scheme we should expect the sequence of the determinants of such values to diminish steadily, according to (50.113). It seems, however, that they fluctuate considerably.

For some further work in this field reference may be made to Bartlett and Rajalakshman (1953) and the monograph by Quenouille (1957), who generalizes the test of 50.9 to the multivariate case.

Systems of equations

50.36 In constructing a mathematical model of a system we are usually led to a specification in terms of a set of relations among various kinds of quantities. For simplicity we shall suppose that these relations are all equations (and not, for example,

inequalities) and that they are linear. In practice this latter condition is not so restrictive as it might appear; sometimes we can get rid of curvilinearities by variate transformations, sometimes a curvilinear relation can be replaced by linear ones in the way that a curve may be replaced approximately by segments of straight lines.

50.37 Outside of the physical sciences, exact mathematical relations of a deterministic kind are rare. In the typical situation we have linear relations among variables which are inexact in the sense that error terms are present. For example, consider the simple assumed relation between two observed variables y and x

$$y = \beta x. \quad (50.117)$$

This may be inexact for at least three reasons: (1) the relationship between y and x is not linear; (2) the observed variables are subject to errors of observation, in which case the true relationship applies to unobservable variables η and ξ ; (3) the relation is exact as far as it goes, but there are other variables also influencing y and the correct relation is

$$y = \beta x + \varepsilon, \quad (50.118)$$

where ε , at this stage, merely stands for something unknown which we cannot specify more explicitly.

50.38 Equation (50.118) is a structural relation among variables which are not necessarily stochastic. It is *not* a regression equation. However, when faced with such relations in practice it is not unreasonable to postulate that ε behaves like a random variable, and to depart from that assumption only when evidence about the actual behaviour of ε is accumulated. We shall, moreover, assume that variables y and x are not subject to errors of observation.

We are thus led to consider systems of equations of linear type which do not incorporate errors of observation but do incorporate a stochastic element. Our object is to use the observations to estimate the constants in these equations and the variances of the stochastic terms. We have already considered some systems of the kind: (a) regressions with independent errors, (b) autoregressions with independent errors, and (c) autoregressions with moving-average errors. We proceed to consider briefly two other types: (d) regressions with autocorrelated errors, and (e) mixed regressive-autoregressive systems.

Regression with autocorrelated errors

50.39 This case appears to have been first discussed in any detail by Cochrane and Orcutt (1949), who pointed out that least-squares estimation was not free from bias when the error terms were correlated. A test for the existence of such correlation was provided by Durbin and Watson (1950-1). Exact results are difficult to obtain, but Durbin and Watson set up a test statistic which, in effect, falls between two other statistics, each of which follows R. L. Anderson's distribution (48.8). See also Watson (1955) and Watson and Hannan (1956).

50.40 Consider a regression of y on fixed x 's,

$$y_t = \beta_1 x_{1t} + \dots + \beta_q x_{qt} + u_t, \quad t = 1, 2, \dots, n, \quad (50.119)$$

where u_t is written instead of the usual ε_t to denote that it is autocorrelated. We may often, without serious error, represent the autocorrelation structure of u by assuming it to be autoregressive:

$$\sum_0^k \alpha_j u_{t-j} = \varepsilon_t. \quad (50.120)$$

If the α 's were known we could transform (50.119) to

$$\sum_{j=0}^k \alpha_j y_{t-j} = \sum_{l=1}^q \beta_l \sum_{j=0}^k \alpha_j x_{l,t-j} + \sum_{j=0}^k \alpha_j u_{t-j}, \quad (50.121)$$

namely to

$$y'_t = \sum_1^q \beta_l x'_t + \varepsilon_t \quad (50.122)$$

where

$$y'_t = \sum_{j=0}^k \alpha_j y_{t-j}, \quad (50.123)$$

$$x'_t = \sum_{j=0}^k \alpha_j x_{l,t-j}. \quad (50.124)$$

Equation (50.122) is now an ordinary regression. Cochrane and Orcutt (1949), to whom this so-called "autoregressive transformation" is due, suggest guessing values of α , estimating β from (50.122), and iterating the process if necessary by recalculating residuals and finding a further approximation to the α 's.

Durbin (1960b) has proposed an alternative procedure which yields asymptotically efficient estimators. Writing $\gamma_{lj} = \beta_l \alpha_j$, we put (50.121) in the form

$$y_t + \sum_{j=1}^k \alpha_j y_{t-j} = \sum_{l,j} \gamma_{lj} x_{l,t-j} + \varepsilon_t. \quad (50.125)$$

If the γ 's were independent we could, as indicated below, regard this asymptotically as a regression of y_t on the other y 's and the x 's, and derive least-squares estimators of α and γ . If the corresponding estimators of α , β , γ are a , b , c we have, in virtue of (50.119),

$$y_t + \sum_{j=1}^k a_j y_{t-j} - \sum c_{lj} x_{l,t-j} = y_t + \sum a_j u_{t-j} - \sum (c_{lj} - a_j \beta_l) x_{l,t-j}. \quad (50.126)$$

Hence the a 's and $(c - a\beta)$'s are least-squares coefficients of regression on u_{t-j} and $x_{l,t-j}$. Consequently the quantities $a_j - \alpha_j$ and $c_{lj} - \alpha_j \beta_l$ are asymptotically normal with zero means and ascertainable dispersion matrix. We can therefore write down their likelihood and maximize it to obtain estimators of α and β .

In certain cases the least-squares estimates derived simply from (50.119) are asymptotically efficient—cf. Grenander and Rosenblatt (1957) and R. L. and T. W. Anderson (1950)—but tests of hypotheses are impaired.

Mixed autoregressive-regressive systems

50.41 Consider now the case where an autoregressive set of y 's is regressed on fixed x 's:

$$\sum_{j=0}^k \alpha_j y_{t-j} = \sum_{l=1}^q \beta_l x_{lt} + \varepsilon_t. \quad (50.127)$$

We can express this in a form similar to (50.126):

$$y_t = - \sum_{j=1}^k \alpha_j y_{t-j} + \sum_{l=1}^q \beta_{ll} x_t + \varepsilon_t. \quad (50.128)$$

However, this is not a regression with fixed variables on the right, owing to the appearance of the lagged y 's. Durbin showed (1960a) that *asymptotically* the properties of least-squares estimators in such a system are the same as those without lagged variables, whether or not the residuals are normally distributed. This is a natural extension of the Mann-Wald theorem mentioned in 50.7.

50.42 We shall not have the space to develop any further the theory of estimation and testing in statistical models, a subject of major importance which is full of pitfalls. Some general comments may, however, be useful.

- (a) It is important to remember which variables are being treated as "fixed" and which are, by their own nature or by the way in which the model is written, stochastic. This is particularly true when equations in these variables are being manipulated. For example, if we denote the random variable by a lower-case letter and a fixed variable by a capital, the regression

$$y = \beta X + \varepsilon \quad (50.129)$$

is not the same thing as

$$x = \frac{1}{\beta} Y - \frac{1}{\beta} \varepsilon. \quad (50.130)$$

- (b) The point becomes of particular interest in time-series wherein the same variable u_t may occur in lagged form u_{t-1} , u_{t-2} , etc. In the equation

$$u_t = \rho u_{t-1} + \varepsilon_t \quad (50.131)$$

we should usually regard both u_t and u_{t-1} as random variables. However, at time t , u_{t-1} has already occurred and is known. It is thus not random in one sense; for example, if (50.131) is regarded as a predictive equation, we are interested in the conditional variable $u_t | u_{t-1}$, not the joint distribution of u_t and u_{t-1} .

- (c) It will be clear, and was forcibly brought to notice by Haavelmo (1943), that estimation of the constants in a subset of equations, instead of the whole set, may result in bias. Thus there is always a further source of error in estimation which must not be forgotten—we may have omitted part of the model.
- (d) The nature of the data available sometimes leads to the specification of incorrect models. For example, the demand for a commodity influences its price, and its price influences supply. But to write

$$\begin{aligned} d_t &= f(p_t) \\ p_t &= g(s_t) \end{aligned} \quad (50.132)$$

overlooks a fundamental property of the system, in that there may be a lag before a change in one variable affects the other. The lag may be so short that its effect does not appear in any statistical evidence we are able to collect; but to ignore it is to destroy the utility of the model.

50.43 A Scandinavian school led by Wold (1964) has insisted on confining economic models to what is known as the "causal-chain" approach, and much of what they have

to say is relevant to the general problem of analysing dynamic systems. The phenomenon under study is conceived of as a chain of causation. A behaviour variable (observable) is subject to causal influences specified by a number of explanatory variables and is influenced by other behaviour variables only through the explanatory set. Theoretically, perhaps, relations expressing the dependence of behaviour variables on explanatory ones should be lagged in time. But when this is not possible the equations are to be regarded as asymmetrical and read from left to right, e.g. the dependence of price on demand, say the simple linear equation

$$p = \alpha d, \quad (50.133)$$

is not invertible to give

$$d = \frac{1}{\alpha} p. \quad (50.134)$$

The literature on model building is scattered, inadequate, and incomplete. That on the statistical analysis of models is worse. A monograph by Fisk (1966) gives a useful account of problems associated with sets of equations. For the causal-chain method see the collection of papers edited by Wold (1964).

Forecasting

50.44 One of the main objects of time-series analysis is to be able to predict the behaviour of the system under study over some future period of time; or, at least, to be able to see whether prediction within acceptable limits of error is possible.

Two approaches to the problem are available. In the first we adopt a purely statistical approach: the past behaviour of a series is studied, and on the assumption that the generating system is constant an attempt is made to project the series into the future without a detailed study of the generating system itself. Thus, given an autoregressive series and having estimated its constants, we may write, for example,

$$u_t = -\hat{\alpha}_1 u_{t-1} - \hat{\alpha}_2 u_{t-2} + \varepsilon_t \quad (50.135)$$

and estimate u_t , (a) by substituting the known values of u_{t-1} and u_{t-2} in this equation, and (b) by assuming that the best estimate we can make of the disturbance term ε is to equate it to zero. If we have, from previous experience, estimates of the variances of ε and of our estimators of α_1 and α_2 , we may put confidence intervals round the estimate of u_t .

50.45 This frankly empirical approach is based on the assumptions (a) that the system is such that an autoregressive scheme (or some other chosen scheme) is a good approximation to the effect of the true generating mechanism, and (b) that such mechanism is not changing, or at any rate not changing rapidly enough to impair the supposition that we may use the equation based on past experience to represent its behaviour in the future. If, however, we wish to delve more deeply into the nature of the generator, we must set up a model; that is to say we must try to write down in specific form the relationships which condition the motion of the system. This is a more complicated exercise, involving on the one hand a much greater insight into the causal mechanisms at work, and on the other hand a lot more effort in estimating the various quantities involved. The tendency has been for statisticians to prefer the simpler approach and

to extrapolate from past experience without attempting to set up a model. This may well be the more rewarding approach for *prediction in the short term*. But it does not enable us to predict what would happen if we altered the system.

50.46 If it has been found that an autoregressive scheme or a scheme of regression satisfactorily fits past experience, there remains little to be said about the forecasting problem. We merely use the authenticated relationship to predict future values. This can be done for any form of time-series. If it has been decomposed into elements such as trend, seasonal, and oscillatory series, we predict the future of each element and reassemble them to forecast the future of the original series. As we remarked at an earlier stage, the underlying supposition is that the various elements are causally independent.

50.47 In practice it is often found that schemes of order two are as satisfactory as such schemes can be, i.e. little is gained by adding extra terms. In fact, a good deal of attention has been given to the case where the scheme is of order one, namely is a Markoff scheme. The prediction equation is then very simple but possibly too simple. A heuristic approach suggested by Holt (1957) has some attractive features.

We consider a scheme of autoregressive type,

$$u_{t+1} = \alpha u_t + \alpha(1-\alpha)u_{t-1} + \alpha(1-\alpha)^2 u_{t-2} + \dots + \alpha(1-\alpha)^k u_{t-k} + \varepsilon_{t+1}. \quad (50.136)$$

Considered as a predictor this has a certain intuitive appeal if $|1-\alpha| < 1$, for then the terms contribute less and less to u_{t+1} as we go back in time. If we estimate u_{t+1} by the systematic component of (50.136), i.e. ignore ε_{t+1} , we have

$$\begin{aligned} \text{Est } u_{t+1} - \text{Est } u_t &= \alpha \sum_0^k (1-\alpha)^j u_{t-j} - \alpha \sum_0^k (1-\alpha)^j u_{t-j-1} \\ &= \alpha(u_t - \text{Est } u_t) - (1-\alpha)^{k+1} u_{t-k-1}. \end{aligned} \quad (50.137)$$

For $|1-\alpha|$ not too close to unity and moderately large k we may write

$$\begin{aligned} \text{Est } u_{t+1} - \text{Est } u_t &= \alpha(u_t - \text{Est } u_t) \\ &= \alpha \varepsilon_t. \end{aligned} \quad (50.138)$$

Suppose α known. At any time-point $t+1$ we know ε_t ; it was the error of estimate at time t . Thus we simply estimate u_{t+1} by taking the estimate at time t and adding $\alpha \varepsilon_t$.

50.48 The estimation of the parameter α is not a simple matter. The most straightforward approach, given enough computational assistance, is to try a range of values of α and to calculate the sum

$$\sum \{u_{t+1} - \alpha u_t - \dots - \alpha(1-\alpha)^k u_{t-k}\}^2$$

for different values of k , selecting the values of α and k which minimize it. In practice it seems that one does not need great precision in the exact determination of optimal α .

Systems of type (50.136) are known, for obvious reasons, as exponentially weighted moving-average predictors. They have been studied in more detail by Brown (1959), Barnard (1959), Cox (1961), Box and Jenkins (1962), and Ward (1963). Winters (1960) proposed an extension which includes seasonal movements.

50.49 At this point we must end, realizing that there are some branches of the subject which might have been discussed at greater length and many more which

remain for future development. The reader who has stayed the course thus far will, we hope, be willing to make allowances for the shortcomings of our work. A subject which is growing as rapidly as ours does not easily shake down into a coherent structure. Nor is this a matter for regret, in that so many of the changes in emphasis are due to growth and an abundant vitality which will undoubtedly carry our subject to further triumphs. We acknowledge our great indebtedness to the writers on whose work we have so freely drawn; we apologize for our errors and omissions; and we write these final words with a considerable sense of relief.

BIBLIOGRAPHICAL NOTE

A comprehensive *Bibliography of Statistical Literature* by M. G. Kendall and Alison G. Doig (Oliver and Boyd, Edinburgh), covering about 30,000 items from the sixteenth century up to 1958, is now available. Volume 1 (1962) covers the years 1950-58, Volume 2 (1965) the years 1940-49, and Volume 3 (1966) the years up to 1939. From 1959 onwards reference should be made to the *Journal of Statistical Abstracts*, published for the International Statistical Institute periodically. There are also in existence a number of specialized bibliographies, including a particularly fine one issued under the editorship of H. Wold (1965).

EXERCISES

50.1 If V_n is the determinant of the matrix of (50.12), show that

$$V_n = (1 + \beta^2)V_{n-1} - \beta^2 V_{n-2}$$

and hence that

$$V_n = (1 - \beta^{2n+2})/(1 - \beta^2).$$

50.2 In the notation of 50.9 show that, for a Markoff scheme,

$$\omega_j = r_j + 2\rho r_{j-1} + \rho^2 r_{j-2},$$

and hence that such a scheme is inadequate to represent the series of Example 50.2.

50.3 For η of equation (50.26) show that if $G(z)$ is the autocovariance function of ε , that of η is

$$\left(\sum_{j=-\infty}^{\infty} \alpha_{k-j} z^j \right) \left(\sum_{j=-\infty}^{\infty} \alpha_{k-j} z^{-j} \right) G(z)$$

and hence that ε and η have the same autocovariance generating function.

50.4 Verify equation (50.32).

50.5 (Progressive solution of the Yule-Walker equations (47.66).) A linear autoregressive scheme is of order k . If the coefficients α are calculated on the assumption that it is of order $s \leq k$, giving $\alpha_{s1}, \alpha_{s2}, \dots, \alpha_{sk}$ and $\alpha_{11} = -\rho_1$, show that

$$\begin{aligned} \alpha_{st} &= \alpha_{s-1,t} + \alpha_{s3} \alpha_{s-1,s-t}, \quad t = 1, 2, \dots, s-1, \\ \alpha_{ss} &= -\frac{\rho_s + \alpha_{s-1,1} \rho_{s-1} + \alpha_{s-2,2} \rho_{s-2} + \dots + \alpha_{s-1,s-1} \rho_1}{1 + \alpha_{s-1,1} \rho_1 + \dots + \alpha_{s-1,s-1} \rho_{s-1}}, \\ &\quad s = 1, 2, \dots, k. \end{aligned}$$

(Durbin, 1960b)

50.6 In the notation of 50.10 show that if the likelihood is written as $(1 - \beta^2)^{-\frac{1}{2}} f(Q) da$, so that

$$\int f(Q) da = (1 - \beta^2)^{\frac{1}{2}},$$

then

$$E\left(\frac{\partial Q}{\partial \beta}\right) = O(n^{-1})$$

$$E\left(\frac{\partial^2 Q}{\partial \beta^2}\right) = \frac{1}{2}nE\left(\frac{\partial Q}{\partial \beta}\right)^2 + O(n^{-1}).$$

Hence show that approximately

$$\text{var } b = E\left(\frac{\partial Q}{\partial \beta}\right)^2 / E\left(\frac{\partial^2 Q}{\partial \beta^2}\right)$$

and derive equation (50.48).

(Durbin, 1959b)

50.7 Verify equation (50.88).

50.8 If two series of linear autoregressive or moving-average type are generated from the same series of random elements, show that for all k

$$\sum_{i=-\infty}^{\infty} \rho_{(11)i} \rho_{(22)k-i} = \sum_{i=-\infty}^{\infty} \rho_{(12)i} \rho_{(21)k-i}.$$

(Quenouille, 1957)

50.9 In Example 50.6, if the matrices \mathbf{F} are the \mathbf{A} -matrices of an autoregressive scheme, show that only one determines a process which is stationary.

50.10 Generally in 50.26, by considering the case where \mathbf{B} is the identity matrix, discuss the conditions under which an autoregressive scheme has an identifiable stationary solution.

Envoi to Volume 3

“Before your going down at the end of the Parliament, I thought good to deliver unto you certain notes for your observation, that serve aptly for the present time, to be imported afterwards when you shall come abroad. . . .

“Yourselves can witness that I never entered into the examination of any cause without advisement, carrying ever a single eye to justice and truth; for, though I were content to hear matters argued and debated pro and contra, as all princes must that will understand what is right, yet I look ever as it were upon a plain table wherein is written neither partiality nor prejudice.”

ELIZABETH I, to her last Parliament

APPENDIX TABLES

- 1 The frequency function of the normal distribution
- 2 The distribution function of the normal distribution
- 3 Quantiles of the d.f. of χ^2
- 4a The distribution function of χ^2 for one degree of freedom, $0 \leq \chi^2 \leq 1$
- 4b The distribution function of χ^2 for one degree of freedom, $1 \leq \chi^2 \leq 10$
- 5 Quantiles of the d.f. of t
- 6 5 per cent points of z
- 7 5 per cent points of F
- 8 1 per cent points of z
- 9 1 per cent points of F
- 10 Symmetric functions. Augmented symmetric functions in terms of power-sums and vice versa

APPENDIX TABLES

506

Appendix Table 1 Frequency function of the normal distribution $y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ with first and second differences

x	y	$\Delta^1 (-)$	Δ^2	x	y	$\Delta^1 (-)$	Δ^2
0.0	0.39894	199	-392	2.5	0.01753	395	+79
0.1	0.39695	591	-374	2.6	0.01358	316	+66
0.2	0.39104	965	-347	2.7	0.01042	250	+53
0.3	0.38139	1312	-308	2.8	0.00792	197	+45
0.4	0.36827	1620	-265	2.9	0.00595	152	+36
0.5	0.35207	1885	-212	3.0	0.00443	116	+27
0.6	0.33322	2097	-159	3.1	0.00327	89	+23
0.7	0.31225	2256	-104	3.2	0.00238	66	+17
0.8	0.28969	2360	-52	3.3	0.00172	49	+13
0.9	0.26609	2412	0	3.4	0.00123	36	+10
1.0	0.24197	2412	+46	3.5	0.00087	26	+7
1.1	0.21785	2366	+84	3.6	0.00061	19	+6
1.2	0.19419	2282	+118	3.7	0.00042	13	+4
1.3	0.17137	2164	+143	3.8	0.00029	9	+2
1.4	0.14973	2021	+161	3.9	0.00020	7	+3
1.5	0.12952	1860	+173	4.0	0.00013	4	—
1.6	0.11092	1687	+177	4.1	0.00009	3	—
1.7	0.09405	1510	+177	4.2	0.00006	2	—
1.8	0.07895	1333	+170	4.3	0.00004	2	—
1.9	0.06562	1163	+162	4.4	0.00002	—	—
2.0	0.05399	1001	+150	4.5	0.00002	—	—
2.1	0.04398	851	+137	4.6	0.00001	—	—
2.2	0.03547	714	+120	4.7	0.00001	—	—
2.3	0.02833	594	+108	4.8	0.00000	—	—
2.4	0.02239	486	+91				

Appendix Table 2 Distribution function of the normal distribution

The table shows the area under the curve $y = (2\pi)^{-1/2} e^{-1/2 x^2}$ lying to the left of specified deviates x ; e.g. the area corresponding to a deviate 1.86 ($= 1.5 + 0.36$) is 0.9686.

Deviate	0.0 +	0.5 +	1.0 +	1.5 +	2.0 +	2.5 +	3.0 +	3.5 +
0.00	5000	6915	8413	9332	9772	9 ² 379	9 ² 865	9 ³ 77
0.01	5040	6950	8438	9345	9778	9 ² 396	9 ² 869	9 ³ 78
0.02	5080	6985	8461	9357	9783	9 ² 413	9 ² 874	9 ³ 78
0.03	5120	7019	8485	9370	9788	9 ² 430	9 ² 878	9 ³ 79
0.04	5160	7054	8508	9382	9793	9 ² 446	9 ² 882	9 ³ 80
0.05	5199	7088	8531	9394	9798	9 ² 461	9 ² 886	9 ³ 81
0.06	5239	7123	8554	9406	9803	9 ² 477	9 ² 889	9 ³ 81
0.07	5279	7157	8577	9418	9808	9 ² 492	9 ² 893	9 ³ 82
0.08	5319	7190	8599	9429	9812	9 ² 506	9 ² 897	9 ³ 83
0.09	5359	7224	8621	9441	9817	9 ² 520	9 ² 900	9 ³ 83
0.10	5398	7257	8643	9452	9821	9 ² 534	9 ³ 03	9 ³ 84
0.11	5438	7291	8665	9463	9826	9 ² 547	9 ³ 06	9 ³ 85
0.12	5478	7324	8686	9474	9830	9 ² 560	9 ³ 10	9 ³ 85
0.13	5517	7357	8708	9484	9834	9 ² 573	9 ³ 13	9 ³ 86
0.14	5557	7389	8729	9495	9838	9 ² 585	9 ³ 16	9 ³ 86
0.15	5596	7422	8749	9505	9842	9 ² 598	9 ³ 18	9 ³ 87
0.16	5636	7454	8770	9515	9846	9 ² 609	9 ³ 21	9 ³ 87
0.17	5675	7486	8790	9525	9850	9 ² 621	9 ³ 24	9 ³ 88
0.18	5714	7517	8810	9535	9854	9 ² 632	9 ³ 26	9 ³ 88
0.19	5753	7549	8830	9545	9857	9 ² 643	9 ³ 29	9 ³ 89
0.20	5793	7580	8849	9554	9861	9 ² 653	9 ³ 31	9 ³ 89
0.21	5832	7611	8869	9564	9864	9 ² 664	9 ³ 34	9 ³ 90
0.22	5871	7642	8888	9573	9868	9 ² 674	9 ³ 36	9 ³ 90
0.23	5910	7673	8907	9582	9871	9 ² 683	9 ³ 38	9 ⁴ 04
0.24	5948	7704	8925	9591	9875	9 ² 693	9 ³ 40	9 ⁴ 08
0.25	5987	7738	8944	9599	9878	9 ² 702	9 ³ 42	9 ⁴ 12
0.26	6026	7764	8962	9608	9881	9 ² 711	9 ³ 44	9 ⁴ 15
0.27	6064	7794	8980	9616	9884	9 ² 720	9 ³ 46	9 ⁴ 18
0.28	6103	7823	8997	9625	9887	9 ² 728	9 ³ 48	9 ⁴ 22
0.29	6141	7852	9015	9633	9890	9 ² 736	9 ³ 50	9 ⁴ 25
0.30	6179	7881	9032	9641	9893	9 ² 744	9 ³ 52	9 ⁴ 28
0.31	6217	7910	9049	9649	9896	9 ² 752	9 ³ 53	9 ⁴ 31
0.32	6255	7939	9066	9656	9898	9 ² 760	9 ³ 55	9 ⁴ 33
0.33	6293	7967	9082	9664	9901	9 ² 767	9 ³ 57	9 ⁴ 36
0.34	6331	7995	9099	9671	9904	9 ² 774	9 ³ 58	9 ⁴ 39
0.35	6368	8023	9115	9678	9906	9 ² 781	9 ³ 60	9 ⁴ 41
0.36	6406	8051	9131	9686	9909	9 ² 788	9 ³ 61	9 ⁴ 43
0.37	6443	8078	9147	9693	9911	9 ² 795	9 ³ 62	9 ⁴ 46
0.38	6480	8106	9162	9699	9913	9 ² 801	9 ³ 64	9 ⁴ 48
0.39	6517	8133	9177	9706	9916	9 ² 807	9 ³ 65	9 ⁴ 50
0.40	6554	8159	9192	9713	9918	9 ² 813	9 ³ 66	9 ⁴ 52
0.41	6591	8186	9207	9719	9920	9 ² 819	9 ³ 68	9 ⁴ 54
0.42	6628	8212	9222	9726	9922	9 ² 825	9 ³ 69	9 ⁴ 56
0.43	6664	8238	9236	9732	9925	9 ² 831	9 ³ 70	9 ⁴ 58
0.44	6700	8264	9251	9738	9927	9 ² 836	9 ³ 71	9 ⁴ 59
0.45	6736	8289	9265	9744	9929	9 ² 841	9 ³ 72	9 ⁴ 61
0.46	6772	8315	9279	9750	9931	9 ² 846	9 ³ 73	9 ⁴ 63
0.47	6808	8340	9292	9756	9932	9 ² 851	9 ³ 74	9 ⁴ 64
0.48	6844	8365	9306	9761	9934	9 ² 856	9 ³ 75	9 ⁴ 66
0.49	6879	8389	9319	9767	9936	9 ² 861	9 ³ 76	9 ⁴ 67

Note—Decimal points in the body of the table are omitted. Repeated 9's are indicated by powers, e.g. 9³71 stands for 0.99971.

APPENDIX TABLES

Appendix Table 3 Quantiles of the d.f. of χ^2
 (Reproduced from Table III of Sir Ronald Fisher's *Statistical Methods for Research Workers*,
 Oliver and Boyd Ltd., Edinburgh, by kind permission of the author and publishers)

$P = 1 - F$	0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
$\nu = 1$	0.003157	0.00628	0.02393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	0.554	0.752	1.145	1.60	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.358	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.821	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

Note—For values of ν greater than 30 the quantity $\sqrt{(2\chi^2)}$ may be taken to be distributed normally about mean $\sqrt{(2\nu - 1)}$ with unit variance

Appendix Table 4a Distribution function of χ^2 for one degree of freedom for values $\chi^2 = 0$ to $\chi^2 = 1$ by steps of 0.01

χ^2	P	Δ	χ^2	P	Δ
0	1.00000	7966	0.50	0.47950	436
0.01	0.92034	3280	0.51	0.47514	430
0.02	0.88754	2505	0.52	0.47084	423
0.03	0.86249	2101	0.53	0.46661	418
0.04	0.84148	1842	0.54	0.46243	411
0.05	0.82306	1656	0.55	0.45832	406
0.06	0.80650	1516	0.56	0.45426	400
0.07	0.79134	1404	0.57	0.45026	395
0.08	0.77730	1312	0.58	0.44631	389
0.09	0.76418	1235	0.59	0.44242	384
0.10	0.75183	1169	0.60	0.43858	379
0.11	0.74014	1111	0.61	0.43479	374
0.12	0.72903	1060	0.62	0.43105	369
0.13	0.71843	1015	0.63	0.42736	365
0.14	0.70828	974	0.64	0.42371	360
0.15	0.69854	938	0.65	0.42011	355
0.16	0.68916	905	0.66	0.41656	351
0.17	0.68011	874	0.67	0.41305	346
0.18	0.67137	845	0.68	0.40959	343
0.19	0.66292	820	0.69	0.40616	338
0.20	0.65472	795	0.70	0.40278	334
0.21	0.64677	773	0.71	0.39944	330
0.22	0.63904	752	0.72	0.39614	326
0.23	0.63152	731	0.73	0.39288	322
0.24	0.62421	713	0.74	0.38966	318
0.25	0.61708	696	0.75	0.38648	315
0.26	0.61012	679	0.76	0.38333	311
0.27	0.60333	663	0.77	0.38022	308
0.28	0.59670	648	0.78	0.37714	304
0.29	0.59022	634	0.79	0.37410	301
0.30	0.58388	620	0.80	0.37109	297
0.31	0.57768	607	0.81	0.36812	294
0.32	0.57161	595	0.82	0.36518	291
0.33	0.56566	583	0.83	0.36227	287
0.34	0.55983	572	0.84	0.35940	285
0.35	0.55411	560	0.85	0.35655	281
0.36	0.54851	551	0.86	0.35374	278
0.37	0.54300	540	0.87	0.35096	276
0.38	0.53760	530	0.88	0.34820	272
0.39	0.53230	521	0.89	0.34548	270
0.40	0.52709	512	0.90	0.34278	267
0.41	0.52197	503	0.91	0.34011	264
0.42	0.51694	495	0.92	0.33747	261
0.43	0.51199	487	0.93	0.33486	258
0.44	0.50712	479	0.94	0.33228	256
0.45	0.50233	471	0.95	0.32972	253
0.46	0.49762	463	0.96	0.32719	251
0.47	0.49299	457	0.97	0.32468	248
0.48	0.48842	449	0.98	0.32220	246
0.49	0.48393	443	0.99	0.31974	243
0.50	0.47950	436	1.00	0.31731	241

APPENDIX TABLES

Appendix Table 4b Distribution function of χ^2 for one degree of freedom for values of χ^2 from 1 to 10 by steps of 0.1

χ^2	P	Δ	χ^2	P	Δ
1.0	0.31731	2304	5.5	0.01902	106
1.1	0.29427	2095	5.6	0.01796	99
1.2	0.27332	1911	5.7	0.01697	94
1.3	0.25421	1749	5.8	0.01603	89
1.4	0.23672	1605	5.9	0.01514	83
1.5	0.22067	1477	6.0	0.01431	79
1.6	0.20590	1361	6.1	0.01352	74
1.7	0.19229	1258	6.2	0.01278	71
1.8	0.17971	1163	6.3	0.01207	66
1.9	0.16808	1078	6.4	0.01141	62
2.0	0.15730	1000	6.5	0.01079	59
2.1	0.14730	929	6.6	0.01020	56
2.2	0.13801	864	6.7	0.00964	52
2.3	0.12937	803	6.8	0.00912	50
2.4	0.12134	749	6.9	0.00862	47
2.5	0.11385	699	7.0	0.00815	44
2.6	0.10686	651	7.1	0.00771	42
2.7	0.10035	609	7.2	0.00729	39
2.8	0.09426	568	7.3	0.00690	38
2.9	0.08858	532	7.4	0.00652	35
3.0	0.08326	497	7.5	0.00617	33
3.1	0.07829	465	7.6	0.00584	32
3.2	0.07364	436	7.7	0.00552	30
3.3	0.06928	408	7.8	0.00522	28
3.4	0.06520	383	7.9	0.00494	26
3.5	0.06137	359	8.0	0.00468	25
3.6	0.05778	337	8.1	0.00443	24
3.7	0.05441	316	8.2	0.00419	23
3.8	0.05125	296	8.3	0.00396	21
3.9	0.04829	279	8.4	0.00375	20
4.0	0.04550	262	8.5	0.00355	19
4.1	0.04288	246	8.6	0.00336	18
4.2	0.04042	231	8.7	0.00318	17
4.3	0.03811	217	8.8	0.00301	16
4.4	0.03594	205	8.9	0.00285	15
4.5	0.03389	192	9.0	0.00270	14
4.6	0.03197	181	9.1	0.00256	14
4.7	0.03016	170	9.2	0.00242	13
4.8	0.02846	160	9.3	0.00229	12
4.9	0.02686	151	9.4	0.00217	12
5.0	0.02535	142	9.5	0.00205	10
5.1	0.02393	134	9.6	0.00195	11
5.2	0.02259	126	9.7	0.00184	10
5.3	0.02133	119	9.8	0.00174	9
5.4	0.02014	112	9.9	0.00165	8
5.5	0.01902	106	10.0	0.00157	8

Appendix Table 5 Quantiles of the d.f. of t
(Reproduced from Sir Ronald Fisher and Dr F. Yates: *Statistical Tables for Biological, Medical and Agricultural Research*,
Oliver and Boyd Ltd., Edinburgh, by kind permission of the authors and publishers)

$P = 2(1 - F)$	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
$\nu = 1$	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.126	0.254	0.387	0.527	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

APPENDIX TABLES

Appendix Table 6 5 per cent points of the distribution of z
(values at which the d.f. = 0.95)

(Reprinted from Table VI of Sir Ronald Fisher's *Statistical Methods for Research Workers*,
Oliver and Boyd Ltd., Edinburgh, by kind permission of the author and publishers)

		Values of ν_1									
		1	2	3	4	5	6	8	12	24	∞
Values of ν_2	1	2.5421	2.6479	2.6870	2.7071	2.7194	2.7276	2.7380	2.7484	2.7588	2.7693
	2	1.4592	1.4722	1.4765	1.4787	1.4800	1.4808	1.4819	1.4830	1.4840	1.4851
	3	1.1577	1.1284	1.1137	1.1051	1.0994	1.0953	1.0899	1.0842	1.0781	1.0716
	4	1.0212	0.9690	0.9429	0.9272	0.9168	0.9093	0.8993	0.8885	0.8767	0.8639
	5	0.9441	0.8777	0.8441	0.8236	0.8097	0.7997	0.7862	0.7714	0.7550	0.7368
	6	0.8948	0.8188	0.7798	0.7558	0.7394	0.7274	0.7112	0.6931	0.6729	0.6499
	7	0.8606	0.7777	0.7347	0.7080	0.6896	0.6761	0.6576	0.6369	0.6134	0.5862
	8	0.8355	0.7475	0.7014	0.6725	0.6525	0.6378	0.6175	0.5945	0.5682	0.5371
	9	0.8163	0.7242	0.6757	0.6450	0.6238	0.6080	0.5862	0.5613	0.5324	0.4979
	10	0.8012	0.7058	0.6553	0.6232	0.6009	0.5843	0.5611	0.5346	0.5035	0.4657
	11	0.7889	0.6909	0.6387	0.6055	0.5822	0.5648	0.5406	0.5126	0.4795	0.4387
	12	0.7788	0.6786	0.6250	0.5907	0.5666	0.5487	0.5234	0.4941	0.4592	0.4156
	13	0.7703	0.6682	0.6134	0.5783	0.5535	0.5350	0.5089	0.4785	0.4419	0.3957
	14	0.7630	0.6594	0.6036	0.5677	0.5423	0.5233	0.4964	0.4649	0.4269	0.3782
	15	0.7568	0.6518	0.5950	0.5585	0.5326	0.5131	0.4855	0.4532	0.4138	0.3628
	16	0.7514	0.6451	0.5876	0.5505	0.5241	0.5042	0.4760	0.4428	0.4022	0.3490
	17	0.7466	0.6393	0.5811	0.5434	0.5166	0.4964	0.4676	0.4337	0.3919	0.3366
	18	0.7424	0.6341	0.5753	0.5371	0.5099	0.4894	0.4602	0.4255	0.3827	0.3253
	19	0.7386	0.6295	0.5701	0.5315	0.5040	0.4832	0.4535	0.4182	0.3743	0.3151
	20	0.7352	0.6254	0.5654	0.5265	0.4986	0.4776	0.4474	0.4116	0.3668	0.3057
	21	0.7322	0.6216	0.5612	0.5219	0.4938	0.4725	0.4420	0.4055	0.3599	0.2971
	22	0.7294	0.6182	0.5574	0.5178	0.4894	0.4679	0.4370	0.4001	0.3536	0.2892
	23	0.7269	0.6151	0.5540	0.5140	0.4854	0.4636	0.4325	0.3950	0.3478	0.2818
	24	0.7246	0.6123	0.5508	0.5106	0.4817	0.4598	0.4283	0.3904	0.3425	0.2749
	25	0.7225	0.6097	0.5478	0.5074	0.4783	0.4562	0.4244	0.3862	0.3376	0.2685
	26	0.7205	0.6073	0.5451	0.5045	0.4752	0.4529	0.4209	0.3823	0.3330	0.2625
	27	0.7187	0.6051	0.5427	0.5017	0.4723	0.4499	0.4176	0.3786	0.3287	0.2569
	28	0.7171	0.6030	0.5403	0.4992	0.4696	0.4471	0.4146	0.3752	0.3248	0.2516
	29	0.7155	0.6011	0.5382	0.4969	0.4671	0.4444	0.4117	0.3720	0.3211	0.2466
	30	0.7141	0.5994	0.5362	0.4947	0.4648	0.4420	0.4090	0.3691	0.3176	0.2419
	60	0.6933	0.5738	0.5073	0.4632	0.4311	0.4064	0.3702	0.3255	0.2654	0.1644
	∞	0.6729	0.5486	0.4787	0.4319	0.3974	0.3706	0.3309	0.2804	0.2085	0

Appendix Table 7 5 per cent points of the variance ratio F
(values at which the d.f. = 0.95)

(Reproduced from Sir Ronald Fisher and Dr F. Yates: *Statistical Tables for Biological, Medical and Agricultural Research*, Oliver and Boyd Ltd., Edinburgh, by kind permission of the authors and publishers)

$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	8	12	24	∞
1	161.40	199.50	215.70	224.60	230.20	234.00	238.90	243.90	249.00	254.30
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

Lower 5 per cent points are found by interchange of ν_1 and ν_2 , i.e. ν_1 must always correspond to the greater mean square.

APPENDIX TABLES

514

Appendix Table 8 1 per cent points of the distribution of z
(values at which the d.f. = 0.99)

(Reprinted from Table VI of Sir Ronald Fisher's *Statistical Methods for Research Workers*,
Oliver and Boyd Ltd., Edinburgh, by kind permission of the author and publishers)

		Values of v_1									
		1	2	3	4	5	6	8	12	24	∞
Values of v_2	1	4.1535	4.2585	4.2974	4.3175	4.3297	4.3379	4.3482	4.3585	4.3689	4.3794
	2	2.2950	2.2976	2.2984	2.2988	2.2991	2.2992	2.2994	2.2997	2.2999	2.3001
	3	1.7649	1.7140	1.6915	1.6786	1.6703	1.6645	1.6569	1.6489	1.6404	1.6314
	4	1.5270	1.4452	1.4075	1.3856	1.3711	1.3609	1.3473	1.3327	1.3170	1.3000
	5	1.3943	1.2929	1.2449	1.2164	1.1974	1.1838	1.1656	1.1457	1.1239	1.0997
	6	1.3103	1.1955	1.1401	1.1068	1.0843	1.0680	1.0460	1.0218	0.9948	0.9643
	7	1.2526	1.1281	1.0672	1.0300	1.0048	0.9864	0.9614	0.9335	0.9020	0.8658
	8	1.2106	1.0787	1.0135	0.9734	0.9459	0.9259	0.8983	0.8673	0.8319	0.7904
	9	1.1786	1.0411	0.9724	0.9299	0.9006	0.8791	0.8494	0.8157	0.7769	0.7305
	10	1.1535	1.0114	0.9399	0.8954	0.8646	0.8419	0.8104	0.7744	0.7324	0.6816
	11	1.1333	0.9874	0.9136	0.8674	0.8354	0.8116	0.7785	0.7405	0.6958	0.6408
	12	1.1166	0.9677	0.8919	0.8443	0.8111	0.7864	0.7520	0.7122	0.6649	0.6061
	13	1.1027	0.9511	0.8737	0.8248	0.7907	0.7652	0.7295	0.6882	0.6386	0.5761
	14	1.0909	0.9370	0.8581	0.8082	0.7732	0.7471	0.7103	0.6675	0.6159	0.5500
	15	1.0807	0.9249	0.8448	0.7939	0.7582	0.7314	0.6937	0.6496	0.5961	0.5269
	16	1.0719	0.9144	0.8331	0.7814	0.7450	0.7177	0.6791	0.6339	0.5786	0.5064
	17	1.0641	0.9051	0.8229	0.7705	0.7335	0.7057	0.6663	0.6199	0.5630	0.4879
	18	1.0572	0.8970	0.8138	0.7607	0.7232	0.6950	0.6549	0.6075	0.5491	0.4712
	19	1.0511	0.8897	0.8057	0.7521	0.7140	0.6854	0.6447	0.5964	0.5366	0.4560
	20	1.0457	0.8831	0.7985	0.7443	0.7058	0.6768	0.6355	0.5864	0.5253	0.4421
	21	1.0408	0.8772	0.7920	0.7372	0.6984	0.6690	0.6272	0.5773	0.5150	0.4294
	22	1.0363	0.8719	0.7860	0.7309	0.6916	0.6620	0.6196	0.5691	0.5056	0.4176
	23	1.0322	0.8670	0.7806	0.7251	0.6855	0.6555	0.6127	0.5615	0.4969	0.4068
	24	1.0285	0.8626	0.7757	0.7197	0.6799	0.6496	0.6064	0.5545	0.4890	0.3967
	25	1.0251	0.8585	0.7712	0.7148	0.6747	0.6442	0.6006	0.5481	0.4816	0.3872
	26	1.0220	0.8548	0.7670	0.7103	0.6699	0.6392	0.5952	0.5422	0.4748	0.3784
	27	1.0191	0.8513	0.7631	0.7062	0.6655	0.6346	0.5902	0.5367	0.4685	0.3701
	28	1.0164	0.8481	0.7595	0.7023	0.6614	0.6303	0.5856	0.5316	0.4626	0.3624
	29	1.0139	0.8451	0.7562	0.6987	0.6576	0.6263	0.5813	0.5269	0.4570	0.3550
	30	1.0116	0.8423	0.7531	0.6954	0.6540	0.6226	0.5773	0.5224	0.4519	0.3481
	60	0.9784	0.8025	0.7086	0.6472	0.6028	0.5687	0.5189	0.4574	0.3746	0.2352
	∞	0.9462	0.7636	0.6651	0.5999	0.5522	0.5152	0.4604	0.3908	0.2913	0

Appendix Table 9 1 per cent points of the variance ratio F
(values at which the d.f. = 0.99)

(Reproduced from Sir Ronald Fisher and Dr F. Yates: *Statistical Tables for Biological, Medical and Agricultural Research*, Oliver and Boyd Ltd., Edinburgh, by kind permission of the authors and publishers)

$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
2	98.49	99.00	99.17	99.25	99.30	99.33	99.36	99.42	99.46	99.50
3	34.12	30.81	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.27	9.89	9.47	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

Lower 1 per cent points are found by interchange of ν_1 and ν_2 , i.e. ν_1 must always correspond to the greater mean square.

APPENDIX TABLES

516

Appendix Table 10 Augmented symmetric functions in terms of power-sums and vice versa
(Reproduced from F. N. David and Kendall (1949) by kind permission of Prof. David and the editors of *Biometrika*)

$$(1) = [1]$$

weight 1
weight 2

	[2]	[1 ²]
(2)	1	-1
(1) ²	1	1

weight 3

	[3]	[21]	[1 ³]
(3)	1	-1	2
(2)(1)	1	1	-3
(1) ³	1	3	1

weight 4

	[4]	[31]	[2 ²]	[21 ²]	[1 ⁴]
(4)	1	-1	-1	2	-6
(3)(1)	1	1	·	-2	8
(2) ²	1	·	1	-1	3
(2)(1) ²	1	2	1	1	-6
(1) ⁴	1	4	3	6	1

weight 5

	[5]	[41]	[32]	[31 ²]	[2 ² 1]	[21 ³]	[1 ⁵]
(5)	1	-1	-1	2	2	-6	24
(4)(1)	1	1	·	-2	-1	6	-30
(3)(2)	1	·	1	-1	-2	5	-20
(3)(1) ²	1	2	1	1	·	-3	20
(2) ² (1)	1	1	2	·	1	-3	15
(2)(1) ³	1	3	4	3	3	1	-10
(1) ⁵	1	5	10	10	15	10	1

weight 6

	[6]	[51]	[42]	[41 ²]	[3 ²]	[321]	[31 ³]	[2 ³]	[2 ² 1 ²]	[21 ⁴]	[1 ⁶]
(6)	1	-1	-1	2	-1	2	-6	2	-6	24	-120
(5)(1)	1	1	·	-2	·	-1	6	·	4	-24	144
(4)(2)	1	·	1	-1	·	-1	3	-3	5	-18	90
(4)(1) ²	1	2	1	1	·	·	-3	·	-1	12	-90
(3) ²	1	·	·	·	1	-1	2	·	2	-8	40
(3)(2)(1)	1	1	1	·	1	1	-3	·	-4	20	-120
(3)(1) ³	1	3	3	3	·	·	1	·	·	-4	40
(2) ³	1	·	3	·	·	3	·	·	·	3	-15
(2) ² (1) ²	1	2	3	1	2	4	·	1	-1	-6	45
(2)(1) ⁴	1	4	7	6	4	16	4	3	1	1	-15
(1) ⁶	1	6	15	15	10	60	20	15	45	15	1

To express the [] functions in terms of (), read downwards up to and including the main diagonal, e.g. [41²] = 2 (6) - 2 (5)(1) - (4)(2) + (4)(1)². To express the () functions in terms of [], read across up to and including the main diagonal, e.g. (4)(1)² = [6] + 2 [51] + [42] + [41²].

REFERENCES

- ANDERSON, O. (1914). Nochmals über "The elimination of spurious correlation due to position in time and space". *Biometrika*, **10**, 269.
- ANDERSON, R. L. (1942). Distribution of the serial correlation coefficient. *Ann. Math. Statist.*, **13**, 1.
- ANDERSON, R. L. and ANDERSON, T. W. (1956). Distribution of the circular serial correlation coefficient for residuals from a fitted Fourier series. *Ann. Math. Statist.*, **21**, 59.
- ANDERSON, R. L. and BANCROFT, T. A. (1952). *Statistical Theory in Research*. McGraw-Hill, New York.
- ANDERSON, T. W. (1946). The non-central Wishart distribution and certain problems of multivariate statistics. *Ann. Math. Statist.*, **17**, 409.
- ANDERSON, T. W. (1958). *Introduction to Multivariate Statistical Analysis*. Wiley, New York.
- ANDERSON, T. W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Ann. Math. Statist.*, **33**, 255.
- ANDERSON, T. W. (1963a). Asymptotic theory for principal component analysis. *Ann. Math. Statist.*, **34**, 122.
- ANDERSON, T. W. (1963b). A test for equality of means when covariance matrices are unequal. *Ann. Math. Statist.*, **34**, 671.
- ANDERSON, T. W. and DAS GUPTA, S. (1964a). Monotonicity of the power functions of some tests of independence between two sets of variates. *Ann. Math. Statist.*, **35**, 206.
- ANDERSON, T. W. and DAS GUPTA, S. (1964b). A monotonicity property of the power functions of some tests of the equality of two covariance matrices. *Ann. Math. Statist.*, **35**, 1059.
- ANDREWS, F. C. (1954). Asymptotic behaviour of some rank tests for analysis of variance. *Ann. Math. Statist.*, **25**, 724.
- ANSCOMBE, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, **35**, 246.
- ANSCOMBE, F. J. (1959). Quick analysis methods for random balance screening experiments. *Technometrics*, **1**, 195.
- ANSCOMBE, F. J. (1961). Examination of residuals. *Proc. 4th Berkeley Symp. Math. Statist. and Prob.* (Univ. California Press), **1**, 1.
- ANSCOMBE, F. J. and TUKEY, J. W. (1963). The examination and analysis of residuals. *Technometrics*, **5**, 141.
- ARMITAGE, P. (1947). A comparison of stratified with unrestricted random sampling from a finite population. *Biometrika*, **34**, 273.
- ARNOLD, H. J. (1964). Permutation support for multivariate techniques. *Biometrika*, **51**, 65.
- ASHTON, E. H., HEALY, M. J. R. and LIPTON, S. (1957). The descriptive use of discriminant functions in physical anthropology. *Proc. Roy. Soc.*, **B**, **146**, 552.

REFERENCES

518

- ATIQULLAH, M. (1962). The estimation of residual variance in quadratically balanced least-squares problems and the robustness of the F -test. *Biometrika*, **94**, 83.
- ATIQULLAH, M. (1964). The robustness of the covariance analysis of a one-way classification. *Biometrika*, **51**, 365.
- BANERJEE, K. S. (1964). A note on idempotent matrices. *Ann. Math. Statist.*, **35**, 880.
- BARNARD, G. A. (1959). Control charts and stochastic processes. *J.R. Statist. Soc.*, **B**, **21**, 239.
- BARNARD, M. M. (1935). The secular variations of skull characters in four series of Egyptian skulls. *Ann. Eugen.*, **6**, 352.
- BARNETT, V. D. and LEWIS, T. (1963). A study of the relations between G.C.E. and degree results. *J.R. Statist. Soc.*, **A**, **126**, 187.
- BARRA, J. R. (1965). Carrés latins et eulériens. *Rev. Int. Statist. Inst.*, **33**, 16.
- BARTHOLOMEW, D. J. (1961). Ordered tests in the analysis of variance. *Biometrika*, **48**, 325.
- BARTLETT, M. S. (1933). On the theory of statistical regression. *Proc. Roy. Soc. Edin.*, **53**, 260.
- BARTLETT, M. S. (1936). The square root transformation in analysis of variance. *Suppl. J.R. Statist. Soc.*, **3**, 68.
- BARTLETT, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Suppl. J.R. Statist. Soc.*, **8**, 27, 85 (Corrigenda, 1948: 10).
- BARTLETT, M. S. (1947a). The use of transformations. *Biometrics*, **3**, 39.
- BARTLETT, M. S. (1947b). The general canonical correlation distribution. *Ann. Math. Statist.*, **18**, 1.
- BARTLETT, M. S. (1947c). Multivariate analysis. *Suppl. J.R. Statist. Soc.*, **9**, 176.
- BARTLETT, M. S. (1950). Periodogram analysis and continuous spectra. *Biometrika*, **37**, 1.
- BARTLETT, M. S. (1951a). The effect of standardization on a χ^2 approximation in factor analysis. *Biometrika*, **38**, 337.
- BARTLETT, M. S. (1951b). The goodness of fit of a single hypothetical discriminant function in the case of several groups. *Ann. Eugen.*, **16**, 199.
- BARTLETT, M. S. (1954). A note on the multiplying factors for various χ^2 approximations. *J.R. Statist. Soc.*, **B**, **16**, 296.
- BARTLETT, M. S. (1955). *An Introduction to Stochastic Processes, with Special Reference to Methods and Applications*. Cambridge Univ. Press.
- BARTLETT, M. S. and KENDALL, D. G. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Suppl. J.R. Statist. Soc.*, **8**, 128.
- BARTLETT, M. S. and RAJALAKSHMAN, D. V. (1953). Goodness of fit tests for simultaneous autoregressive series. *J.R. Statist. Soc.*, **B**, **15**, 107.
- BASU, D. (1958). On sampling with and without replacement. *Sankhyā*, **20**, 287.
- BENARD, A. and VAN ELTEREN, P. (1953). A generalization of the method of m rankings. *Proc. Kon. Ned. Akad. Wetensch.*, **A**, **56** (*Indag. Math.*, **15**, 358).

- BENNETT, B. M. (1951). Note on the solution of the generalized Behrens-Fisher problem. *Ann. Inst. Statist. Math., Tokyo*, **2**, 87.
- BEVERIDGE, W. H. (1921). Weather and harvest cycles. *Econ. J.*, **31**, 429.
- BHAPKAR, V. P. (1963). The asymptotic power and efficiency of Mood's test for two-way classification. *J. Indian Statist. Ass.*, **1**, 24.
- BHUCHONGKUL, S. and PURI, M. L. (1965). On the estimation of contrasts in linear models. *Ann. Math. Statist.*, **36**, 198.
- BICKEL, P. J. (1965). On some robust estimates of location. *Ann. Math. Statist.*, **36**, 847.
- BIGGERS, J. D. (1959). The estimation of missing and mixed-up observations in several experimental designs. *Biometrika*, **46**, 91.
- BIRKHOFF, G. D. (1931). Proof of the ergodic theorem. *Proc. Nat. Acad. Sci., U.S.A.*, **17**, 656.
- BLACKITH, R. E. (1960). A synthesis of multivariate techniques to distinguish patterns of growth in grasshoppers. *Biometrics*, **16**, 28.
- BLACKMAN, R. B. and TUKEY, J. W. (1958). *The Measurement of Power Spectra from the Point of View of Communications Engineering*. Dover Co., New York.
- BOSE, R. C. (1936). On the exact distribution and moment-coefficients of the D^2 statistic. *Sankhyā*, **2**, 143.
- BOSE, R. C. (1939). On the construction of balanced incomplete block designs. *Ann. Eugen.*, **9**, 353.
- BOSE, R. C. (1942). A note on the resolvability of balanced incomplete block designs. *Sankhyā*, **6**, 105.
- BOSE, R. C. (1956). Paired comparison designs for testing concordance between judges. *Biometrika*, **43**, 113.
- BOSE, R. C. and CARTER, R. L. (1959). Complete representation in the construction of rotatable designs. *Ann. Math. Statist.*, **30**, 771.
- BOSE, R. C., CLATWORTHY, W. H. and SHRIKAND, S. S. (1954). Tables of partially balanced designs with two associate classes. *North Carolina Agric. Exper. Sta. Tech. Bull. no. 107*.
- BOSE, R. C. and DRAPER, N. R. (1959). Second order rotatable designs in three dimensions. *Ann. Math. Statist.*, **30**, 1097.
- BOSE, R. C. and NAIR, K. R. (1939). Partially balanced incomplete block designs. *Sankhyā*, **4**, 337.
- BOSE, R. C. and ROY, S. N. (1938). The distribution of the studentized D^2 statistic. *Sankhyā*, **4**, 19.
- BOSE, R. C. and SHRIKAND, S. S. (1959). On the falsity of Euler's conjecture about the non-existence of two orthogonal latin squares of order $4t + 2$. *Proc. Nat. Acad. Sci., U.S.A.*, **45**, 734.
- BOSE, R. C. and SHRIKAND, S. S. (1960). On the construction of sets of mutually orthogonal latin squares and the falsity of a conjecture of Euler. *Trans. Amer. Math. Soc.*, **95**, 191.
- BOSE, R. C., SHRIKAND, S. S. and PARKER, R. (1960). Further results on the construction of mutually orthogonal latin squares and the falsity of Euler's conjecture. *Canadian J. Math.*, **12**, 189.

- Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, **36**, 317.
- Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification. II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Ann. Math. Statist.*, **25**, 290, 484.
- Box, G. E. P. (1957). Evolutionary operation: a method for increasing industrial productivity. *Appl. Statist.*, **6**, 81.
- Box, G. E. P. and ANDERSEN, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *J.R. Statist. Soc.*, **B**, **17**, 1.
- Box, G. E. P. and BEHNKEN, D. W. (1960). Simplex-sum designs: a class of second order rotatable designs derivable from those of first order. *Ann. Math. Statist.*, **31**, 838.
- Box, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J.R. Statist. Soc.*, **B**, **26**, 211.
- Box, G. E. P. and HUNTER, J. S. (1957). Multi-factor experimental designs for exploring response surfaces. *Ann. Math. Statist.*, **28**, 195.
- Box, G. E. P. and JENKINS, G. M. (1962). Some statistical aspects of adaptive optimization and control. *J.R. Statist. Soc.*, **B**, **24**, 297.
- Box, G. E. P. and TIDWELL, P. W. (1962). Transformation of the independent variables. *Technometrics*, **4**, 531.
- Box, G. E. P. and WATSON, G. S. (1962). Robustness to non-normality of regression tests. *Biometrika*, **49**, 93.
- Box, G. E. P. and WILSON, K. B. (1951). On the experimental attainment of optimum conditions. *J.R. Statist. Soc.*, **B**, **13**, 1.
- BOZIVICH, H., BANCROFT, T. A. and HARTLEY, H. O. (1956). Power of analysis of variance test procedures for certain incompletely specified models, I. *Ann. Math. Statist.*, **27**, 1017.
- BRADU, D. (1965). Main effect analysis of the general non-orthogonal layout with any number of factors. *Ann. Math. Statist.*, **36**, 88.
- BRILLINGER, D. R. (1965). An introduction to polyspectra. *Ann. Math. Statist.*, **36**, 1351.
- BROWN, G. W. and MOOD, A. M. (1951). On median tests for linear hypotheses. *Proc. 2nd Berkeley Symp. Math. Statist. and Prob.* (Univ. California Press), 159.
- BROWN, R. G. (1959). *Statistical Forecasting for Inventory Control*. McGraw-Hill, New York.
- BUDNE, T. A. (1959). The application of random balance designs. *Technometrics*, **1**, 139.
- BULMER, M. G. (1957). Approximate confidence limits for components of variance. *Biometrika*, **44**, 159.
- BURMAN, J. P. (1965). Moving seasonal adjustment of economic time series. *J.R. Statist. Soc.*, **A**, **128**, 534; also same *J.* (1966), **129**, 274.
- CHACKO, V. J. (1963). Testing homogeneity against ordered alternatives. *Ann. Math. Statist.*, **34**, 945.

- COCHRAN, W. G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. *Ann. Eugen.*, **11**, 47.
- COCHRAN, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, **3**, 22.
- COCHRAN, W. G. (1961). Comparison of methods for determining stratum boundaries. *Bull. Int. Statist. Inst.*, **38** (2), 345.
- COCHRAN, W. G. (1963). *Sampling Techniques*. 2nd edn. Wiley, New York.
- COCHRAN, W. G. (1964). On the performance of the linear discriminant function. *Technometrics*, **6**, 179.
- COCHRAN, W. G. (1965). The planning of observational studies of human populations. *J.R. Statist. Soc., A*, **128**, 234.
- COCHRAN, W. G. and BLISS, C. I. (1948). Discriminant functions with covariance. *Ann. Math. Statist.*, **19**, 151.
- COCHRAN, W. G. and COX, G. M. (1957). *Experimental Designs*. 2nd edn. Wiley, New York.
- COCHRAN, W. G. and HOPKINS, C. E. (1961). Some classification problems with multivariate qualitative data. *Biometrics*, **17**, 10.
- COCHRANE, D. and ORCUTT, G. H. (1949). Application of least-squares regression to relationships containing auto-correlated error terms. *J. Amer. Statist. Ass.*, **44**, 32.
- COLLIER, R. O. JR. and BAKER, F. B. (1966). Some Monte Carlo results on the power of the *F*-test under permutation in the simple randomized block design. *Biometrika*, **53**, 199.
- CORNFIELD, J. and TUKEY, J. W. (1956). Average values of mean squares in factorials. *Ann. Math. Statist.*, **27**, 907.
- COWDEN, D. J. (1962). Weights for fitting polynomial secular trends. *Univ. North Carolina Sch. Business Admin. Tech. Paper no. 4*.
- COX, D. R. (1952). Sequential tests for composite hypotheses. *Proc. Camb. Phil. Soc.*, **48**, 290.
- COX, D. R. (1958a). *Planning of Experiments*. John Wiley, New York.
- COX, D. R. (1958b). The interpretation of the effects of non-additivity in the latin square. *Biometrika*, **45**, 69.
- COX, D. R. (1961). Prediction by exponentially weighted moving averages and related methods. *J.R. Statist. Soc., B*, **23**, 414.
- CRADDOCK, J. M. (1965). A meteorological application of principal component analysis. *The Statistician*, **15**, 143.
- CURTISS, J. H. (1943). On transformations used in the analysis of variance. *Ann. Math. Statist.*, **14**, 107.
- DALENIUS, T. (1950). The problem of optimum stratification. *Skand. Aktuartskr.*, **33**, 203.
- DALENIUS, T. (1952). The problem of optimum stratification in a special type of design. *Skand. Aktuartskr.*, **35**, 61.
- DALENIUS, T. (1953). The economics of one-stage stratified sampling. *Sankhyā*, **12**, 351.

- DALENIUS, T. and HODGES, J. L., Jr. (1957, 1959). The choice of stratification points. *Skand. Aktuarietidskr.*, **40**, 198. Minimum variance stratification. *J. Amer. Statist. Ass.*, **54**, 88.
- DALY, J. F. (1940). On the unbiased character of likelihood-ratio tests for independence in normal systems. *Ann. Math. Statist.*, **11**, 1.
- DANIELL, P. J. (1946). Discussion on "Symposium on autocorrelation in time series." *Suppl. J.R. Statist. Soc.*, **8**, 88.
- DANIELS, H. E. (1956). The approximate distribution of serial correlation coefficients. *Biometrika*, **43**, 169.
- DANIELS, H. E. (1962). The estimation of spectral densities. *J.R. Statist. Soc., B*, **24**, 185.
- DARLING, D. A. (1952). On a test for homogeneity and extreme values. *Ann. Math. Statist.*, **23**, 450 (correction: **24**, 135).
- DARROCH, J. N. and SILVEY, S. D. (1963). On testing more than one hypothesis. *Ann. Math. Statist.*, **34**, 555.
- DAS GUPTA, S., ANDERSON, T. W. and MUDHOLKAR, G. S. (1964). Monotonicity of the power functions of some tests of the multivariate linear hypothesis. *Ann. Math. Statist.*, **35**, 200.
- DAVID, H. A. (1963). *The Method of Paired Comparisons*. Griffin, London.
- DAVID, H. A. and ARENS, B. E. (1959). Optimal spacing in regression analysis. *Ann. Math. Statist.*, **30**, 1072.
- DAVIES, O. L. (also editor), BOX, G. E. P., CONNOR, L. R., COUSINS, W. R., HIMS-WORTH, F. R. and SILLITO, G. P. (1954). *The Design and Analysis of Industrial Experiments*. Oliver and Boyd, Edinburgh.
- DAVIS, H. T. (1941). The analysis of economic time series. *Cowles Comm. Res. Econ.* (Bloomington, Indiana), Monogr. no. 6.
- DEMPSTER, A. P. (1960-1). Random allocation designs. I. On general classes of estimation methods. II. Approximate theory for simple random allocation. *Ann. Math. Statist.*, **31**, 885; **32**, 387.
- DEMPSTER, A. P. (1966). Estimation in multivariate analysis. *Proc. Symp. Multiv. Analysis, Dayton, Ohio*.
- DIXON, W. J. (1944). Further contributions to the problem of serial correlation. *Ann. Math. Statist.*, **15**, 119.
- DOLBY, J. L. (1963). A quick method for choosing a transformation. *Technometrics*, **5**, 317.
- DRAPER, N. R. (1960a). Second order rotatable designs in four or more dimensions. *Ann. Math. Statist.*, **31**, 23.
- DRAPER, N. R. (1960b). Third order rotatable designs in three dimensions. *Ann. Math. Statist.*, **31**, 865.
- DRAPER, N. R. (1960c). A third order rotatable design in four dimensions. *Ann. Math. Statist.*, **31**, 875.
- DRAPER, N. R. (1961). Third order rotatable designs in three dimensions: some specific designs. *Ann. Math. Statist.*, **32**, 910.
- DUNCAN, D. B. (1952). On the properties of the multiple comparisons test. *Virginia J. Sci., N.S.*, **3**, 49.

- DUNCAN, D. B. (1955). Multiple range and multiple F tests. *Biometrics*, **11**, 1.
- DUNCAN, D. B. (1957). Multiple range tests for correlated and heteroscedastic means. *Biometrics*, **13**, 164.
- DUNN, O. J. (1961). Multiple comparisons among means. *J. Amer. Statist. Ass.*, **56**, 52.
- DURBIN, J. (1951). Incomplete blocks in ranking experiments. *Brit. J. Psychol. (Statist. Sec.)*, **4**, 85.
- DURBIN, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. *J.R. Statist. Soc.*, **B**, **15**, 262.
- DURBIN, J. (1954). Non-response and call-backs in surveys. *Bull. Int. Statist. Inst.*, **34** (2), 72.
- DURBIN, J. (1958). Sampling theory for estimates based on fewer individuals than the number selected. *Bull. Int. Statist. Inst.*, **36** (3), 113.
- DURBIN, J. (1959a). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, **46**, 477.
- DURBIN, J. (1959b). Efficient estimation of parameters in moving-average models. *Biometrika*, **46**, 306.
- DURBIN, J. (1960a). Estimation of parameters in time series regression models. *J.R. Statist. Soc.*, **B**, **22**, 139.
- DURBIN, J. (1960b). The fitting of time-series models. *Rev. Int. Statist. Inst.*, **28**, 233.
- DURBIN, J. (1961). Efficient fitting of linear models for continuous stationary time series from discrete data. *Bull. Int. Statist. Inst.*, **38** (4), 273.
- DURBIN, J. (1965). A method of sample selection with unequal probabilities without replacement. (Abstract.) *Ann. Math. Statist.*, **36**, 1327.
- DURBIN, J. and WATSON, G. S. (1950-1). Testing for serial correlation in least squares regression. I, II. *Biometrika*, **37**, 409; **38**, 159.
- EISENHART, C. (1947). The assumptions underlying the analysis of variance. *Biometrics*, **3**, 1.
- EISENPRESS, H. (1956). Regression techniques applied to seasonal corrections and adjustments for calendar shifts. *J. Amer. Statist. Ass.*, **51**, 615.
- EKMAN, G. (1959). An approximation useful in univariate stratification. *Ann. Math. Statist.*, **30**, 219.
- FELLEGI, I. P. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. *J. Amer. Statist. Ass.*, **58**, 183.
- FINCH, P. D. (1960). On the covariance determinants of moving-average and autoregressive models. *Biometrika*, **47**, 194.
- FINNEY, D. J. (1941). The joint distribution of variance ratios based on a common error mean square. *Ann. Eugen.*, **11**, 136.
- FINNEY, D. J. (1952). *Statistical Method in Biological Assay*. Griffin, London.
- FISHER, R. A. (1929). Tests of significance in harmonic analysis. *Proc. Roy. Soc.*, **A**, **125**, 54.
- FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

- FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179.
- FISHER, R. A. (1939). The sampling distribution of some statistics obtained from non-linear equations. *Ann. Eugen.*, **9**, 238.
- FISHER, R. A. (1942). The theory of confounding in factorial experiments in relation to the theory of groups. *Ann. Eugen.*, **11**, 341.
- FISHER, R. A. (1945). A system of confounding for factors with more than two alternatives, giving completely orthogonal cubes and higher powers. *Ann. Eugen.*, **12**, 283.
- FORTIER, J. J. and SOLOMON, H. (1966). Clustering procedures. *Proc. Symp. Multiv. Analysis*, Dayton, Ohio.
- FOSTER, F. G. (1957-8). Upper percentage points of the generalized Beta distribution. II, III. *Biometrika*, **44**, 441; **45**, 492.
- FOSTER, F. G. and REES, D. H. (1957). Upper percentage points of the generalized Beta distribution. I. *Biometrika*, **44**, 237.
- FOSTER, F. G. and STUART, A. (1954). Distribution-free tests in time-series based on the breaking of records. *J.R. Statist. Soc.*, **B**, **16**, 1.
- FREEMAN, G. H. and JEFFERS, J. N. R. (1962). Estimation of means and standard errors in the analysis of non-orthogonal experiments by electronic computer. *J.R. Statist. Soc.*, **B**, **24**, 435.
- FREEMAN, M. F. and TUKEY, J. W. (1950). Transformations related to the angular and the square root. *Ann. Math. Statist.*, **21**, 607.
- FRIEDMAN, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Ass.*, **32**, 675.
- FRIEDMAN, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Statist.*, **11**, 86.
- GABRIEL, K. R. (1963). Analysis of variance of proportions with unequal frequencies. *J. Amer. Statist. Ass.*, **58**, 1133.
- GABRIEL, K. R. (1964). A procedure for testing the homogeneity of all sets of means in analysis of variance. *Biometrics*, **20**, 459.
- GABRIEL, K. R. (1966). Simultaneous test procedures with applications. *Ann. Math. Statist.*, to be published.
- GARDINER, D. A., GRANDAGE, A. H. E. and HADER, R. J. (1959). Third order rotatable designs for exploring response surfaces. *Ann. Math. Statist.*, **30**, 1082.
- GASSNER, B. J. (1965). Equal-difference BIB designs. *Proc. Amer. Math. Soc.*, **16**, 378.
- GAUTSCHI, W. (1959). Some remarks on Herbach's paper, "Optimum nature of the F -test for Model II in the balanced case." *Ann. Math. Statist.*, **30**, 960.
- GAYLOR, D. W. and SWEENEY, H. C. (1965). Design for optimal prediction in simple linear regression. *J. Amer. Statist. Ass.*, **60**, 205.
- GEARY, R. C. (1944). Extension of a theorem by Harald Cramér on the frequency distribution of the quotient of two variables. *J.R. Statist. Soc.*, **107**, 56.
- GEISSER, S. (1964). Estimation in the uniform covariance case. *J.R. Statist. Soc.*, **B**, **26**, 477.

- GHOSH, B. K. (1964). Simultaneous tests by sequential methods in hierarchical classifications. *Biometrika*, **51**, 439.
- GHOSH, M. N. and SHARMA, D. (1963). Power of Tukey's test for non-additivity. *J.R. Statist. Soc.*, **B**, **25**, 213.
- GHOSH, S. P. (1963a). Post cluster sampling. *Ann. Math. Statist.*, **34**, 587.
- GHOSH, S. P. (1963b). Optimum stratification with two characters. *Ann. Math. Statist.*, **34**, 866.
- GIRSHICK, M. A. (1939). On the sampling theory of the roots of determinantal equations. *Ann. Math. Statist.*, **10**, 203.
- GLEISSBERG, W. (1945). Eine Aufgabe der Kombinatorik und Wahrscheinlichkeitsrechnung. *Univ. Istanbul Rev. Fac. Sci.*, **A**, **10**, 25.
- GLESER, L. J. (1966). A note on the sphericity test. *Ann. Math. Statist.*, **37**, 464.
- GODAMBE, V. P. (1955). A unified theory of sampling from finite populations. *J.R. Statist. Soc.*, **B**, **17**, 269.
- GODAMBE, V. P. (1965). A review of the contributions towards a unified theory of sampling from finite populations. *Rev. Int. Statist. Inst.*, **33**, 242.
- GOOD, I. J. (1950). The inversion of circulant matrices. *Biometrika*, **37**, 185.
- GOODMAN, L. A. and HARTLEY, H. O. (1958). The precision of unbiased ratio-type estimators. *J. Amer. Statist. Ass.*, **53**, 491.
- GRANGER, C. W. J. (1963). The effect of varying month length on the analysis of economic time series. *L'industrie*, **1**, 41.
- GRANGER, C. W. J. and HATANAKA, M. (1964). *Spectral Analysis of Economic Time Series*. Princeton Univ. Press, Princeton.
- GRAYBILL, F. A. and HULTQUIST, R. A. (1961). Theorems concerning Eisenhart's Model II. *Ann. Math. Statist.*, **32**, 261.
- GRAYBILL, F. A. and MARSAGLIA, G. (1957). Idempotent matrices and quadratic forms in the general linear hypothesis. *Ann. Math. Statist.*, **28**, 678.
- GRAYBILL, F. A. and SESHADRI, V. (1960). On the unbiasedness of Yates' method of estimation using interblock information. *Ann. Math. Statist.*, **31**, 786.
- GRAYBILL, F. A. and WEEKS, D. L. (1959). Combining inter-block and intra-block information in balanced incomplete blocks. *Ann. Math. Statist.*, **30**, 799.
- GRENANDER, U. and ROSENBLATT, M. (1957). *Statistical Analysis of Stationary Time Series*. Wiley, New York.
- GRUNDY, P. M. (1954). A method of sampling with probability exactly proportional to size. *J.R. Statist. Soc.*, **B**, **16**, 236.
- GUÉRIN, R. (1965). Vue d'ensemble sur les plans en blocs incomplets équilibrés et partiellement équilibrés. *Rev. Int. Statist. Inst.*, **33**, 24.
- GUEST, P. G. (1958). The spacing of observations in polynomial regression. *Ann. Math. Statist.*, **29**, 294.
- HAAVELMO, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, **11**, 1.
- HAJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Pub. Math. Inst. Hung. Acad. Sci.*, **B**, **5**, 361.

- HAJEK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.*, **35**, 1491.
- HANANI, H. (1961). The existence and construction of balanced incomplete block designs. *Ann. Math. Statist.*, **32**, 361.
- HANNAN, E. J. (1960). The estimation of seasonal variation. *Australian J. Statist.*, **2**, 1.
- HANNAN, E. J. (1961). Testing for a jump in the spectral function. *J.R. Statist. Soc.*, **B**, **23**, 394.
- HANSEN, M. H. and HURWITZ, W. N. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.*, **14**, 333.
- HANSEN, M. H. and HURWITZ, W. N. (1946). The problem of non-response in sample surveys. *J. Amer. Statist. Ass.*, **41**, 517.
- HANSEN, M. H. and HURWITZ, W. N. (1949). On the determination of optimum probabilities in sampling. *Ann. Math. Statist.*, **20**, 426.
- HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953). *Sample Survey Methods and Theory*. 2 vols. Wiley, New York.
- HARMAN, H. H. (1960). *Modern Factor Analysis*. Univ. Chicago Press.
- HARTER, H. L. (1957). Error rates and sample sizes for range tests in multiple comparisons. *Biometrics*, **13**, 511.
- HARTLEY, H. O. (1938). Studentization and large-sample theory. *Suppl. J.R. Statist. Soc.*, **5**, 80.
- HARTLEY, H. O. (1955). Some recent developments in analysis of variance. *Commun. Pure Appl. Math.*, **8**, 47.
- HARTLEY, H. O. (1959). Analytic studies of survey data in Istituto di Statistica (Roma), *Volume in Onore di Corrado Gini*, **1**, 213.
- HARTLEY, H. O. and ROSS, A. (1954). Unbiased ratio estimators. *Nature*, **174**, 270.
- HEALY, M. J. R. and TAYLOR, L. R. (1962). Tables for power-law transformations. *Biometrika*, **49**, 557.
- HENDERSON, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, **9**, 226.
- HERBACH, L. H. (1959). Properties of Model II-type analysis of variance tests, A: optimum nature of the F -test for Model II in the balanced case. *Ann. Math. Statist.*, **30**, 939.
- HESS, I., SETHI, V. K. and BALAKRISHNAN, T. R. (1966). Stratification: a practical investigation. *J. Amer. Statist. Ass.*, **61**, 74.
- HEXT, G. (1964). A note on pre-whitening and recolouring. *Stanford Univ. Dept. Statist. Tech. Rep. no. 13*.
- HIGHAM, J. A. (1882). On the adjustment of mortality tables. *J. Inst. Actuar.*, **23**, 335.
- HIGHAM, J. A. (1885). On the graduation of mortality tables. *J. Inst. Actuar.*, **25**, 245.
- HILLS, M. (1966). Allocation rules and their error rates. *J.R. Statist. Soc.*, **B**, **28**, 1.
- HODGES, J. L., Jr. and LEHMANN, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *Ann. Math. Statist.*, **33**, 482.

- HODGES, J. L., Jr. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.*, **34**, 598.
- HOEL, P. G. (1937). A significance test for component analysis. *Ann. Math. Statist.*, **8**, 149.
- HOEL, P. G. (1958). Efficiency problems in polynomial estimation. *Ann. Math. Statist.*, **29**, 1134.
- HOEL, P. G. (1965a). Minimax designs in two dimensional regression. *Ann. Math. Statist.*, **36**, 1097.
- HOEL, P. G. (1965b). Optimum designs for polynomial extrapolation. *Ann. Math. Statist.*, **36**, 1483.
- HOEL, P. G. and LEVINE, A. (1964). Optimal spacing and weighting in polynomial prediction. *Ann. Math. Statist.*, **35**, 1553.
- HOGG, R. V. (1961). On the resolution of statistical hypotheses. *J. Amer. Statist. Ass.*, **56**, 978.
- HOLT, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. *Carnegie Inst. Technol. Res. Memo. no. 52 (NONR 760(01))*.
- HORSNELL, G. (1953). The effect of unequal group variances on the F -test for the homogeneity of group means. *Biometrika*, **40**, 128.
- HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Ass.*, **47**, 663.
- HOTELLING, H. (1931). The generalization of Student's ratio. *Ann. Math. Statist.*, **2**, 360.
- HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika*, **28**, 321.
- HOTELLING, H. (1951). A generalized T test and measure of multivariate dispersion. *Proc. 2nd Berkeley Symp. Math. Statist. and Prob.*, 23.
- HOWE, W. G. (1955). Some contributions to factor analysis. *U.S. Atom. Energy Comm. Rep. ORNL-1919*.
- HOYLAND, A. (1965). Robustness of the Hodges-Lehmann estimates for shift. *Ann. Math. Statist.*, **36**, 174.
- HSU, P. L. (1939). On the distribution of roots of certain determinantal equations. *Ann. Eugen.*, **9**, 250.
- IMHOF, J. P. (1960). A mixed model for the complete three-way layout with two random-effects factors. *Ann. Math. Statist.*, **31**, 906.
- ITO, K. (1962). A comparison of the powers of two multivariate analysis of variance tests. *Biometrika*, **49**, 455.
- ITO, K. and SCHULL, W. J. (1964). On the robustness of the T_0^2 test in multivariate analysis of variance when variance-covariance matrices are not equal. *Biometrika*, **51**, 71.
- JAMES, A. T. (1964). Distributions of matrix variates and latent roots derived from normal samples. *Ann. Math. Statist.*, **35**, 475.
- JAMES, A. T. (1966). Inference on latent roots by calculation of hypergeometric functions of matrix argument. *Proc. Symp. Multiv. Analysis, Dayton, Ohio*.

- JENKINS, G. M. (1954-6). Tests on hypotheses in the linear autoregressive model. I: Null hypothesis distributions in the Yule scheme. II: Null distributions for higher order schemes; non-null distributions. *Biometrika*, **41**, 405; **43**, 186.
- JENKINS, G. M. (1961). General considerations in the analysis of spectra. Comments on the discussions of Messrs Tukey and Goodman. *Technometrics*, **3**, 133, 229.
- JOHN, S. (1961). Errors in discrimination. *Ann. Math. Statist.*, **32**, 1125.
- JOHNSON, N. L. (1953). Some notes on the application of sequential methods in the analysis of variance. *Ann. Math. Statist.*, **24**, 614.
- JOHNSON, N. L. (1954). Sequential procedures in certain component of variance problems. *Ann. Math. Statist.*, **25**, 357.
- JOHNSON, N. L. (1957). Optimal sampling for quota fulfilment. *Biometrika*, **44**, 518.
- KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments*. Wiley, New York.
- KENDALL, M. G. (1947). The estimation of parameters in linear autoregressive time series. *Econometrica Suppl.*, **17**, 44.
- KENDALL, M. G. (1954). Note on bias in the estimation of autocorrelation. *Biometrika*, **41**, 403.
- KENDALL, M. G. (1957). The moments of the Leipnik distribution. *Biometrika*, **44**, 270.
- KENDALL, M. G. (1961a). A theorem in trend analysis. *Biometrika*, **48**, 224.
- KENDALL, M. G. (1961b). *A Course in Multivariate Analysis*, 2nd imp. Griffin, London.
- KENDALL, M. G. and SMITH, B. BABINGTON (1939). The problem of m rankings. *Ann. Math. Statist.*, **10**, 275.
- KEULS, M. (1952). The use of "Studentized range" in connection with an analysis of variance. *Euphytica*, **1**, 112.
- KHINTCHIN, A. YA (1932). Zu Birkhoff's Lösung des Ergodenproblems. *Math. Ann.*, **107**, 485.
- KIEFER, J. (1958). On the nonrandomized optimality and randomized nonoptimality of symmetrical designs. *Ann. Math. Statist.*, **29**, 675.
- KIEFER, J. (1959). Optimum experimental designs. *J.R. Statist. Soc.*, **B**, **21**, 272.
- KIEFER, J. and WOLFOWITZ, J. (1954). Optimum designs in regression problems. *Ann. Math. Statist.*, **30**, 271.
- KISH, L. and HESS, I. (1959). On variances of ratios and their differences in multi-stage samples. *J. Amer. Statist. Ass.*, **54**, 416.
- KOOP, J. C. (1963). On the axioms of sample formation and their bearing on the construction of linear estimators in sampling theory for finite universes. I, II, III. *Metrika*, **7**, 81 and 165.
- KOOPMANS, T. C. (1942). Serial correlation and quadratic forms in normal variables. *Ann. Math. Statist.*, **13**, 14.
- KRUSKAL, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *J.R. Statist. Soc.*, **B**, **27**, 251.
- KSHIRSAGAR, A. M. (1959). Bartlett decomposition and Wishart distribution. *Ann. Math. Statist.*, **30**, 239.

- KSHIRSAGAR, A. M. (1961). The non-central multivariate beta distribution. *Ann. Math. Statist.*, **32**, 104.
- LAHIRI, D. B. (1951). A method of sample selection providing unbiased ratio estimates. *Bull. Int. Statist. Inst.*, **33** (2), 133.
- LAWLEY, D. N. (1939). A generalization of Fisher's z -test. *Biometrika*, **30**, 180.
- LAWLEY, D. N. (1956a). Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika*, **43**, 128.
- LAWLEY, D. N. (1956b). A general method of approximating to the distribution of likelihood ratio criteria. *Biometrika*, **43**, 295.
- LAWLEY, D. N. (1959). Tests of significance in canonical analysis. *Biometrika*, **46**, 59.
- LAWLEY, D. N. (1963). On testing a set of correlation coefficients for equality. *Ann. Math. Statist.*, **34**, 149.
- LAWLEY, D. N. and MAXWELL, A. E. (1963). *Factor Analysis as a Statistical Method*. Butterworths, London.
- LEDERMANN, W. (1937). On the rank of the reduced correlational matrix in multiple-factor analysis. *Psychometrika*, **2**, 85.
- LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.
- LEHMANN, E. L. (1963a). Robust estimation in analysis of variance. *Ann. Math. Statist.*, **34**, 957.
- LEHMANN, E. L. (1963b). Asymptotically nonparametric inference: an alternative approach to linear models. *Ann. Math. Statist.*, **34**, 1494.
- LEHMANN, E. L. (1963c). Nonparametric confidence intervals for a shift parameter. *Ann. Math. Statist.*, **34**, 1507.
- LEHMANN, E. L. (1964). Asymptotically nonparametric inference in some linear models with one observation per cell. *Ann. Math. Statist.*, **35**, 726.
- LEIPNIK, R. B. (1947). Distribution of the serial correlation coefficient in a circularly correlated universe. *Ann. Math. Statist.*, **18**, 80.
- LEIPNIK, R. B. (1958). Note on the characteristic function of a serial-correlation distribution. *Biometrika*, **45**, 559.
- LEVENE, H. (1952). On the power function of tests of randomness based on runs up and down. *Ann. Math. Statist.*, **23**, 34.
- LINHART, H. (1959). Techniques for discriminant analysis with discrete variables. *Metrika*, **2**, 138.
- LOMNICKI, Z. A. (1961). Tests for departure from normality in the case of linear stochastic processes. *Metrika*, **4**, 37.
- LOW, L. Y. (1964). Sampling variances of estimates of components of variance from a non-orthogonal two-way classification. *Biometrika*, **51**, 491.
- LUBISCHEW, A. A. (1962). On the use of discriminant functions in taxonomy. *Biometrics*, **18**, 455.
- MADOW, W. G. (1945). Note on the distribution of the serial correlation coefficient. *Ann. Math. Statist.*, **16**, 308.

REFERENCES

- 530
- MADOW, W. G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Ann. Math. Statist.*, **19**, 535.
- MAHALANOBIS, P. C. (1930). On tests and measures of group divergence. I. *J. Proc. Asiat. Soc. Bengal*, **26**, 541.
- MANLEY, G. (1953 and later). The mean temperature of Central England, 1698-1952. *Quart. J.R. Meteor. Soc.*, **79**, 242.
- MANN, H. B. (1945a). On a test for randomness based on signs of differences. *Ann. Math. Statist.*, **16**, 193.
- MANN, H. B. (1945b). Nonparametric tests against trend. *Econometrica*, **13**, 245.
- MANN, H. B. (1949). *Analysis and Design of Experiments*. Dover, New York.
- MANN, H. B. and WALD, A. (1943). On the statistical treatment of linear stochastic difference equations. *Econometrica*, **11**, 173.
- MARRIOTT, F. H. C. and POPE, J. A. (1954). Bias in the estimation of autocorrelations. *Biometrika*, **41**, 390.
- MAUCHLY, J. W. (1940). Significance test for sphericity of a normal n -variate distribution. *Ann. Math. Statist.*, **11**, 204.
- MAULDON, J. G. (1955). Pivotal quantities for Wishart's and related distributions, and a paradox in fiducial theory. *J.R. Statist. Soc.*, **B**, **17**, 79.
- MICKEY, M. R. (1959). Some finite population unbiased ratio and regression estimators. *J. Amer. Statist. Ass.*, **54**, 594.
- MIKHAIL, N. N. (1965). A comparison of tests of the Wilks-Lawley hypothesis in multivariate analysis. *Biometrika*, **52**, 149.
- MILL, J. S. (1843). *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation*. 2 vols, London.
- MOOD, A. M. (1951). On the distribution of the characteristic roots of normal second-moment matrices. *Ann. Math. Statist.*, **22**, 266.
- MOORE, G. H. and WALLIS, W. A. (1943). Time series significance tests based on signs of differences. *J. Amer. Statist. Ass.*, **38**, 153.
- MORAN, P. A. P. (1947-8). Some theorems on time series. I, II. The significance of the serial correlation coefficient. *Biometrika*, **34**, 281; **35**, 255.
- MORAN, P. A. P. (1949). The spectral theory of discrete stochastic processes. *Biometrika*, **36**, 63.
- MULLER, E.-R. (1965). A method of constructing balanced incomplete block designs. *Biometrika*, **52**, 285.
- MURTEIRA, B. (1951). Note on the variate differences of autoregressive series. *Biometrika*, **38**, 479.
- MURTHY, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, **18**, 379.
- MURTHY, M. N. (1963). Some recent advances in sampling theory. *J. Amer. Statist. Ass.*, **58**, 737.
- MURTHY, M. N. and NANJAMMA, N. S. (1959). Almost unbiased ratio estimates based on interpenetrating sub-sample estimates. *Sankhyā*, **21**, 381.
- MURTY, V. N. (1961). An inequality for balanced incomplete block designs. *Ann. Math. Statist.*, **32**, 908.

- NANJAMMA, N. S., MURTHY, M. N. and SETHI, V. K. (1959). Some sampling systems providing unbiased ratio estimators. *Sankhyā*, **21**, 299.
- NARAIN, R. D. (1950). On the completely unbiased character of tests of independence in multivariate normal systems. *Ann. Math. Statist.*, **21**, 293.
- NERLOVE, M. (1964). Spectral analysis of seasonal adjustment procedures. *Econometrica*, **32**, 241.
- NEWMAN, D. (1939). The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika*, **31**, 20.
- NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *J. Amer. Statist. Ass.*, **33**, 101.
- NEYMAN, J. and SCOTT, E. L. (1960). Correction for bias introduced by a transformation of variables. *Ann. Math. Statist.*, **31**, 643.
- NIETO DE PASCUAL, J. (1961). Unbiased ratio estimators in stratified sampling. *J. Amer. Statist. Ass.*, **56**, 70.
- NOETHER, G. E. (1950). Asymptotic properties of the Wald-Wolfowitz test of randomness. *Ann. Math. Statist.*, **21**, 231.
- NORTON, H. W. (1939). The 7×7 squares. *Ann. Eugen.*, **9**, 269 (errata: **10**, 1).
- OGAWA, J. (1961). The effect of randomization on the analysis of randomized block design. *Ann. Inst. Statist. Math., Tokyo*, **13**, 105.
- OGAWA, J. (1963). On the null-distribution of the F -statistic in a randomized balanced incomplete block design under the Neyman model. *Ann. Math. Statist.*, **34**, 1558.
- OLKIN, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika*, **45**, 154.
- PARZEN, E. (1961). Mathematical considerations in the estimation of spectra. Comments on the discussions of Messrs Tukey and Goodman. *Technometrics*, **3**, 167, 232.
- PATHAK, P. K. (1961a). Use of "order-statistic" in sampling without replacement. *Sankhyā*, **A**, **23**, 409.
- PATHAK, P. K. (1961b). On the evaluation of moments of distinct units in a sample. *Sankhyā*, **A**, **23**, 415.
- PATHAK, P. K. (1962a). On simple random sampling with replacement. *Sankhyā*, **A**, **24**, 287.
- PATHAK, P. K. (1962b). On sampling with unequal probabilities. *Sankhyā*, **A**, **24**, 315.
- PATHAK, P. K. (1964a). On sampling schemes providing unbiased ratio estimators. *Ann. Math. Statist.*, **35**, 222.
- PATHAK, P. K. (1964b). On inverse sampling with unequal probabilities. *Biometrika*, **51**, 185.
- PATHAK, P. K. (1964c). Sufficiency in sampling theory. *Ann. Math. Statist.*, **35**, 795.
- PATTERSON, H. D. (1950). Sampling on successive occasions with partial replacement of units. *J.R. Statist. Soc.*, **B**, **12**, 241.

REFERENCES

532

- PEARCE, S. C. (1963). The use and classification of non-orthogonal designs. *J.R. Statist. Soc., A*, 126, 353.
- PEARSON, E. S. and WILKS, S. S. (1933). Methods of statistical analysis appropriate for k samples of two variables. *Biometrika*, 25, 353.
- PHILLIPS, A. W. (1959). The estimation of parameters in systems of stochastic differential equations. *Biometrika*, 46, 67.
- PILLAI, K. C. S. (1956). On the distribution of the largest or the smallest root of a matrix in multivariate analysis. *Biometrika*, 43, 122.
- PILLAI, K. C. S. (1964). On the distribution of the largest of seven roots of a matrix in multivariate analysis. *Biometrika*, 51, 270.
- PILLAI, K. C. S. (1966). On the non-central multivariate Beta distribution and the moments of traces of some matrices. *Proc. Symp. Multiv. Analysis, Dayton, Ohio*.
- PITMAN, E. J. G. (1938). Significance tests which may be applied to samples from any population. III. The analysis of variance test. *Biometrika*, 29, 322.
- PLACKETT, R. L. (1950). Some theorems in least squares. *Biometrika*, 37, 149.
- PLACKETT, R. L. (1960). Models in the analysis of variance. *J.R. Statist. Soc., B*, 22, 195.
- POSTEN, H. O. and BARGMANN, R. E. (1964). Power of the likelihood-ratio test of the general linear hypothesis in multivariate analysis. *Biometrika*, 51, 467.
- QUENOUILLE, M. H. (1947a). Notes on the calculation of autocorrelations of linear autoregressive schemes. *Biometrika*, 34, 365.
- QUENOUILLE, M. H. (1947b). A large-sample test for the goodness of fit of autoregressive schemes. *J.R. Statist. Soc.*, 110, 123.
- QUENOUILLE, M. H. (1948). Some results in the testing of serial correlation coefficients. *Biometrika*, 35, 261.
- QUENOUILLE, M. G. (1949a). A method of trend elimination. *Biometrika*, 36, 75.
- QUENOUILLE, M. H. (1949b). The joint distribution of serial correlation coefficients. *Ann. Math. Statist.*, 20, 561.
- QUENOUILLE, M. H. (1953a). *The Design and Analysis of Experiment*. Griffin, London.
- QUENOUILLE, M. H. (1953b). Modifications to the variate-difference method. *Biometrika*, 40, 383.
- QUENOUILLE, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353.
- QUENOUILLE, M. H. (1957). *The Analysis of Multiple Time-Series*. Griffin, London.
- QUENOUILLE, M. H. (1958). Discrete autoregressive schemes with varying time-intervals. *Metrika*, 1, 21.
- RAJ, D. (1956). Some estimators in sampling with varying probabilities without replacement. *J. Amer. Statist. Ass.*, 51, 269.
- RAJ, D. (1964). On double sampling for PPS estimation. *Ann. Math. Statist.*, 35, 900.
- RAJ, D. and KHAMIS, S. H. (1958). Some remarks on sampling with replacement. *Ann. Math. Statist.*, 29, 550.
- RAO, C. R. (1947). General methods of analysis for incomplete block designs. *J. Amer. Statist. Ass.*, 42, 541.

- RAO, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. Wiley, New York.
- RAO, C. R. (1961). A study of BIB designs with replications 11 to 15. *Sankhyā*, A, 23, 117.
- RAO, C. R. and SLATER, P. (1949). Multivariate analysis applied to differences between neurotic groups. *Brit. J. Psychol. (Statist. Sec.)*, 2, 17.
- RAO, J. N. K. (1963). On three procedures of unequal probability sampling without replacement. *J. Amer. Statist. Ass.*, 58, 202.
- RAO, J. N. K. (1965). A note on estimation of ratios by Quenouille's method. *Biometrika*, 52, 647.
- RAO, J. N. K., HARTLEY, H. O. and COCHRAN, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *J.R. Statist. Soc.*, B, 24, 482.
- REES, D. H. (1966). The analysis of variance of designs with many non-orthogonal classifications. *J.R. Statist. Soc.*, B, 28, 110.
- RHODES, E. C. (1921). Smoothing. *Tracts for Computers*, no. 6, 1. Cambridge Univ. Press.
- ROBSON, D. S. (1957). Applications of multivariate polykays to the theory of unbiased ratio-type estimation. *J. Amer. Statist. Ass.*, 52, 511.
- ROBSON, D. S. and VITHAYASAI, C. (1961). Unbiased componentwise ratio estimation. *J. Amer. Statist. Ass.*, 56, 350.
- ROMANOVSKY, V. I. (1932). Sur la loi sinusoidale limite. *R.C. Circ. Mat. Palermo*, 56, 82.
- ROMANOVSKY, V. I. (1933). Sur une généralisation de la loi sinusoidale limite. *R.C. Circ. Mat. Palermo*, 57, 130.
- ROSENBLATT, M. (ed.) (1963). *Proceedings of the Symposium on Time Series Analysis held at Brown University, June 11-14, 1962*. Wiley, New York.
- ROSS, A. (1961). Variance estimates in "optimum" sample designs. *J. Amer. Statist. Ass.*, 56, 135.
- ROY, J. and SHAH, K. R. (1962). Recovery of interblock information. *Sankhyā*, A, 24, 269.
- ROY, S. N. (1939). p -statistics or some generalizations in analysis of variance appropriate to multivariate problems. *Sankhyā*, 4, 381.
- ROY, S. N. (1957). *Some Aspects of Multivariate Analysis*. Wiley, New York.
- ROY, S. N. and COBB, W. (1960). Mixed model variance analysis with normal error and possibly non-normal other random effects. Part I. The univariate case. *Ann. Math. Statist.*, 31, 939.
- RUBIN, H. (1945). On the distribution of the serial correlation coefficient. *Ann. Math. Statist.*, 16, 211.
- SAMPFORD, M. R. (1962). Methods of cluster sampling with and without replacement for clusters of unequal sizes. *Biometrika*, 49, 27.
- SATTERTHWAITE, F. E. (1941). Synthesis of variance. *Psychometrika*, 6, 309.
- SATTERTHWAITE, F. E. (1959). Random balance experimentation. *Technometrics*, 1, 111.

REFERENCES

534

- SCHATZOFF, M. (1966). Sensitivity comparison among tests of the general linear hypothesis. *J. Amer. Statist. Ass.*, **61**, 415.
- SCHEFFÉ, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, **40**, 87.
- SCHEFFÉ, H. (1956a). A "mixed model" for the analysis of variance. *Ann. Math. Statist.*, **27**, 23.
- SCHEFFÉ, H. (1956b). Alternative models for the analysis of variance. *Ann. Math. Statist.*, **27**, 251.
- SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York.
- SCHMETTERER, L. (1960). On a problem of J. Neyman and E. Scott. *Ann. Math. Statist.*, **31**, 656.
- SEARLE, S. R. (1958). Sampling variances of estimates of components of variance. *Ann. Math. Statist.*, **29**, 167.
- SEBER, G. A. F. (1964a). Orthogonality in analysis of variance. *Ann. Math. Statist.*, **35**, 705.
- SEBER, G. A. F. (1964b). Linear hypotheses and induced tests. *Biometrika*, **51**, 41.
- SEN, A. R. (1953). On the estimate of variance in sampling with varying probabilities. *J. Indian. Soc. Agric. Statist.*, **5**, 119.
- SHAH, K. R. (1964). Use of inter-block information to obtain uniformly better estimators. *Ann. Math. Statist.*, **35**, 1064.
- SHISKIN, J. (1955). Seasonal computations on Univac. *Amer. Statistician*, **9** (1), 19.
- SHISKIN, J. and EISENPRESS, H. (1957). Seasonal adjustments by electronic computer methods. *J. Amer. Statist. Ass.*, **52**, 415.
- SIMAICA, J. B. (1941). On an optimum property of two important statistical tests. *Biometrika*, **32**, 70.
- SIOTANI, M. (1964). Interval estimation for linear combinations of means. *J. Amer. Statist. Ass.*, **59**, 1141.
- SITGREAVES, R. (1952). On the distribution of two random matrices used in classification procedures. *Ann. Math. Statist.*, **23**, 263.
- SLUTZKY, E. (1937). The summation of random causes as the source of cyclic processes. *Econometrica*, **5**, 105.
- SMITH, C. A. B. (1947). Some examples of discrimination. *Ann. Eugen.*, **13**, 272 (errata: **14**, opp. p. 279).
- SMITH, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants, and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, **12**, 1.
- SNEDECOR, G. W. (1946). *Statistical methods: applied to experiments in agriculture and biology*. 4th edn. Iowa State College Collegiate Press, Ames, Iowa.
- SPROTT, D. A. (1956). A note on combined interblock and intrablock estimation in incomplete block designs. *Ann. Math. Statist.*, **27**, 633 (correction: **28**, 269).
- SPROTT, D. A. (1962). Listing of BIB designs from $r = 16$ to 20. *Sankhyā*, **A**, **24**, 203.
- SRIVASTAVA, S. R. and BOZIVICH, H. (1962). Power of certain analysis of variance test procedures involving preliminary tests. *Bull. Int. Statist. Inst.*, **39** (3), 133.

- STEVENS, W. L. (1948). Statistical analysis of a non-orthogonal tri-factorial experiment. *Biometrika*, **35**, 346.
- STUART, A. (1952). The power of two difference-sign tests. *J. Amer. Statist. Ass.*, **47**, 416.
- STUART, A. (1954). Asymptotic relative efficiencies of distribution-free tests of randomness against normal alternatives. *J. Amer. Statist. Ass.*, **49**, 147.
- STUART, A. (1956). The efficiencies of tests of randomness against normal regression. *J. Amer. Statist. Ass.*, **51**, 285.
- STUART, A. (1957). The efficiency of the records test for trend in normal regression. *J.R. Statist. Soc.*, **B**, **19**, 149.
- STUART, A. (1964). Multistage sampling with preliminary random stratification of first-stage units. *Rev. Int. Statist. Inst.*, **32**, 193.
- "STUDENT" (1908). The probable error of a mean. *Biometrika*, **6**, 1.
- TAMURA, R. (1966). Multivariate nonparametric several-sample tests. *Ann. Math. Statist.*, **37**, 611.
- THOMPSON, W. A., Jr. (1962). The problem of negative estimates of variance components. *Ann. Math. Statist.*, **33**, 273.
- TIN, M. (1965). Comparison of some ratio estimators. *J. Amer. Statist. Ass.*, **60**, 294.
- TINTNER, G. (1940). The variate difference method. *Cowles Comm. Res. Econ. (Bloomington, Indiana) Monogr. no. 5*.
- TOCHER, K. D. (1952). The design and analysis of block experiments. *J.R. Statist. Soc.*, **B**, **14**, 45.
- TRYON, R. C. (1939). *Cluster Analysis*. Edwards Bros., Ann Arbor, Michigan.
- TUKEY, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, **5**, 232.
- TUKEY, J. W. (1951). Quick and dirty methods in statistics. II. Simple analyses for standard designs. *Proc. 5th Annu. Conf. Amer. Soc. Qual. Contr.*, 189.
- TUKEY, J. W. (1952). Allowances for various types of error rates. Unpublished, Princeton University.
- TUKEY, J. W. (1953). The problem of multiple comparisons. Unpublished, Princeton University.
- TUKEY, J. W. (1956-7). Variances of variance components. I. Balanced designs. II. The unbalanced single classification. *Ann. Math. Statist.*, **27**, 722; **28**, 43.
- TUKEY, J. W. (1957b). On the comparative anatomy of transformations. *Ann. Math. Statist.*, **28**, 602.
- VAN ELTEREN, P. and NOETHER, G. E. (1959). The asymptotic efficiency of the χ_r^2 -test for a balanced incomplete block design. *Biometrika*, **46**, 475.
- VOS, J. W. E. (1964). Sampling in space and time. *Rev. Int. Statist. Inst.*, **32**, 226.
- WAGLE, B. V. (1962). *Some Contributions to the Theory of Multivariate Analysis*. Unpublished Ph.D. Thesis, Univ. London.
- WALD, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *Ann. Math. Statist.*, **15**, 145.

REFERENCES

- 536
- WALD, A. and WOLFOWITZ, J. (1943). An exact test for randomness in the non-parametric case based on serial correlation. *Ann. Math. Statist.*, **14**, 378.
- WALKER, A. M. (1961). On Durbin's formula for the limiting generalized variance of a sample of consecutive observations from a moving-average process. *Biometrika*, **48**, 197 (corrigenda: 476).
- WALKER, G. T. (1914). Correlation in seasonal variation of weather. III. On the criterion for the reality of relationships or periodicities. *Indian Meteor. Dept. (Simla) Mem.*, **21** (9), 22.
- WALKER, G. T. (1931). On periodicity in series of related terms. *Proc. Roy. Soc., A*, **131**, 518.
- WALLIS, W. A. and MOORE, G. H. (1941). A significance test for time series analysis. *J. Amer. Statist. Ass.*, **36**, 401.
- WARD, D. H. (1963). Comparison of different systems of exponentially weighted prediction. *The Statistician*, **13**, 173.
- WATSON, G. S. (1955). Serial correlation in regression analysis. I. *Biometrika*, **42**, 327.
- WATSON, G. S. (1956). On the joint distribution of the circular serial correlation coefficients. *Biometrika*, **43**, 161.
- WATSON, G. S. and HANNAN, E. J. (1956). Serial correlation in regression analysis. II. *Biometrika*, **43**, 436.
- WELCH, B. L. (1937). On the z -test in randomized blocks and latin squares. *Biometrika*, **29**, 21.
- WHITE, J. S. (1957). Approximate moments for the serial correlation coefficient. *Ann. Math. Statist.*, **28**, 798.
- WHITE, J. S. (1961). Asymptotic expansions for the mean and variance of the serial correlation coefficient. *Biometrika*, **48**, 85.
- WHITTAKER, E. T. (1911). On the law which governs the variations of S.S. Cygni. *Month. Notes Roy. Astron. Soc.*, **71**, 686.
- WHITTLE, P. (1951). *Hypothesis Testing in Time Series Analysis*. Almqvist & Wiksell, Uppsala.
- WHITTLE, P. (1953a). The analysis of multiple stationary time series. *J.R. Statist. Soc., B*, **15**, 125.
- WHITTLE, P. (1953b). Estimation and information in stationary time-series. *Ark. Mat. (Stockholm)*, **2**, 423.
- WHITTLE, P. (1957). Curve and periodogram smoothing. *J.R. Statist. Soc., B*, **19**, 38.
- WIENER, N. (1930). Generalized harmonic analysis. *Acta Math.*, **55**, 117.
- WIJSMAN, R. A. (1957). Random orthogonal transformations and their use in some classical distribution problems in multivariate analysis. *Ann. Math. Statist.*, **28**, 415.
- WILK, M. B. and KEMPTHORNE, O. (1955). Fixed, mixed and random models. *J. Amer. Statist. Ass.*, **50**, 1144.
- WILK, M. B. and KEMPTHORNE, O. (1956). Some aspects of the analysis of factorial experiments in a completely randomized design. *Ann. Math. Statist.*, **27**, 950.
- WILK, M. B. and KEMPTHORNE, O. (1957). Non-additivities in a latin square design. *J. Amer. Statist. Ass.*, **52**, 218.

- WILKINSON, G. N. (1957). The analysis of covariance with incomplete data. *Biometrics*, **13**, 363.
- WILKINSON, G. N. (1958a). Estimation of missing values for the analysis of incomplete data. *Biometrics*, **14**, 257.
- WILKINSON, G. N. (1958b). The analysis of variance and derivation of standard errors for incomplete data. *Biometrics*, **14**, 360.
- WILKINSON, G. N. (1960). Comparison of missing value procedures. *Australian J. Statist.*, **2**, 53.
- WILKS, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, **24**, 471.
- WILKS, S. S. (1935). On the independence of k sets of normally distributed statistical variables. *Econometrica*, **3**, 309.
- WILLIAMS, E. J. (1952). Some exact tests in multivariate analysis. *Biometrika*, **39**, 17.
- WILLIAMS, W. H. (1961). Generating unbiased ratio and regression estimators. *Biometrics*, **17**, 267.
- WILLIAMS, W. H. (1962). On two methods of unbiased estimation with auxiliary variates. *J. Amer. Statist. Ass.*, **57**, 184.
- WINTERS, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Sci.*, **6**, 324.
- WISE, J. (1955). The autocorrelation function and the spectral density function. *Biometrika*, **42**, 151.
- WISE, J. (1956). Stationarity conditions for stochastic processes of the autoregressive and moving-average type. *Biometrika*, **43**, 215.
- WISHART, J. (1928). The generalized product moment distribution in samples from a normal multivariate population. *Biometrika*, **20A**, 32, 424.
- WISHART, J. (1929). The correlation between product moments of any order in samples from a normal population. *Proc. Roy. Soc. Edin.*, **49**, 78.
- WISHART, J. (1948). Proofs of the distribution law of the second order moment statistics. *Biometrika*, **35**, 55, 422.
- WISHART, J. and BARTLETT, M. S. (1932). The distribution of second order moment statistics in a normal system. *Proc. Camb. Phil. Soc.*, **28**, 455.
- WISHART, J. and BARTLETT, M. S. (1933). The generalized product moment distribution in a normal system. *Proc. Camb. Phil. Soc.*, **29**, 260.
- WOLD, H. O. A. (1938). *A Study in the Analysis of Stationary Time Series*. 2nd edn, 1954. Almqvist & Wiksell, Stockholm.
- WOLD, H. O. A. (1949). A large-sample test for moving averages. *J.R. Statist. Soc.*, **B**, **11**, 297.
- WOLD, H. O. A. (ed.) (1964). *Econometric Model Building: Essays on the Causal Chain Approach*. North Holland Publishing Co., Amsterdam.
- WOLD, H. O. A. (1965). *Bibliography on Time Series and Stochastic Processes: an International Team Project*. Oliver & Boyd, Edinburgh.
- WOLFOWITZ, J. (1944). Asymptotic distribution of runs up and down. *Ann. Math. Statist.*, **15**, 163.
- WORKING, H. (1960). Note on the correlation of first differences of averages in a random chain. *Econometrica*, **28**, 916.

REFERENCES

538

- YATES, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire J. Exper. Agric.*, **1**, 129.
- YATES, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *J. Amer. Statist. Ass.*, **29**, 51.
- YATES, F. (1936a). Incomplete randomized blocks. *Ann. Eugen.*, **7**, 121.
- YATES, F. (1936b). A new method of arranging variety trials involving a large number of varieties. *J. Agric. Sci.*, **26**, 424.
- YATES, F. (1937). The design and analysis of factorial experiments. *Imp. Bur. Soil Sci. (Harpenden) Tech. Comm. no. 35*.
- YATES, F. (1939). The recovery of inter-block information in variety trials arranged in three-dimensional lattices. *Ann. Eugen.*, **9**, 136.
- YATES, F. (1940a). The recovery of inter-block information in balanced incomplete block designs. *Ann. Eugen.*, **10**, 317.
- YATES, F. (1940b). Lattice squares. *J. Agric. Sci.*, **30**, 672.
- YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, 3rd edn. Griffin, London.
- YATES, F. and GRUNDY, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *J.R. Statist. Soc.*, **B**, **15**, 253.
- YOUNDEN, W. J., KEMPTHORNE, O., TUKEY, J. W., BOX, G. E. P. and HUNTER, J. S. (1959). Discussion of the papers of Messrs Satterthwaite and Budne. *Technometrics*, **1**, 157 (authors' response to discussions: 184).
- YOUNG, D. H. (1961). Quota fulfilment using unrestricted random sampling. *Biometrika*, **48**, 333.
- YULE, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series?—A study in sampling and the nature of time-series. *J.R. Statist. Soc.*, **89**, 1.
- YULE, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Phil. Trans.*, **A**, **226**, 267.

INDEX

(References are to pages)

- Additivity, 13; Tukey's test for, (Example 35.3) 23, 131-2; transformations to, 94.
- Analysis of variance (AV), in linear model (Model I), 1-56; definitions, 2, 6, 57; geometrical interpretations, 9-10, 40; main effects and interactions, 12, 36; choice of weights, 25-6; empty cells, 30; combination of tests, 40-3; multiple comparisons, 43-9; orthogonal polynomials, 49; ordered alternatives, 49; covariance analysis, 49-52; in Model II, 57-75, 76, 78; unbiased quadratic estimation, 59-60; sufficient statistics, 60-2, (Exercise 36.12) 83; correspondence with Model I AV, 62-3; testing hypotheses in Model II, 66-8, 69-70, (Exercises 36.1-6, 36.8), 82-3; power, confidence intervals, negative estimates in Model II, 71, (Exercises 36.11, 36.15), 83-4; unbalanced case of Model II, 72-3; generalization of models, 74-9; mixed models, 77-9; sequential, 79; allocation of units, randomization, 79-81; choice of model, 81-2; transformations, 85-96; analysis of residuals, 96-7; robustness, 97-105; distribution-free methods, 105-9; median tests, 109-11; missing observations, 111-13; for block experiments, 128-9; for randomized blocks, 131; for Latin Squares, 134, (Exercise 38.6) 162; for BIB, 146; interblock information, 149; in factorial experiments, 155; *see* Classification, one-way, etc.
- Andersen, S. L., robustness of AV, 98.
- Anderson, O., variate-difference method, 388, 390.
- Anderson, R. L., Yates' method of weighted squares of means, 30; orthogonal polynomials in AV, 49; analysis of covariance, 52; distribution of serial correlations, 437-40; autocorrelated errors, 498.
- Anderson, T. W., testing degree of polynomial regression, (Exercise 37.3) 114; non-central Wishart distribution, 259; distribution of generalized variance, (Exercise 41.9) 261; tests of independence, 271; multivariate analysis, 281; power functions of multivariate tests, 281; multivariate regression, (Exercises 42.15-16) 283-4; latent roots, 254; discrimination, (Exercises 44.2-3) 339-40; autocorrelated errors, 498.
- Andrews, F. C., median test, 109.
- Angular transformations, 95 (Exercises 37.4-5) 114.
- Anscombe, F. J., transformations, 89, (Exercises 37.4-5) 114-15; analysis of residuals, 96-7, (Exercises 37.15-16) 117; random balance experiments, 130.
- Arens, B. E., polynomial regression designs, 161.
- Armitage, P., stratified sampling, 180, 182, (Exercises 39.14-15) 207.
- Arnold, H. J., T^2 test under permutations, 281.
- Ashton, E. H., data on male premolars, (Example 42.4) 280-1
- Atiqullah, M., robustness of AV, 105.
- Autocorrelation, 404; in ergodic processes, 410; spectrum as g.f. for, 410, 412; function for continuous series, 422; partial, 424-5; bias in estimating, 433-5; of errors in regression, 497-8; *see* Serial correlation.
- Autocovariance, 404; spectrum as g.f. for, 412; *see* Autocorrelation, Time-series.
- Autoregressive series, 416-21; and moving averages, 417, 474-6, 481-4; Yule-Walker equations, 417, (Exercise 50.5) 502; variance, (Exercise 47.7) 427; representable as compound Markoff-Yule series, (Exercise 47.12) 427; matrix representation, (Exercises 47.16-17) 428-9; large-sample theory of serial correlations, 433; variate-differences, (Exercises 48.15-16) 452; effect of starting-point on estimation, 472-4; estimation, 476-8; partial autocorrelations, 478-80; test of fit, 480-1; with moving-average errors, 485-6; mixed-regressive systems, 498-500; *see* Markoff series, Yule series.
- Baker, F. B., robustness of F -test in randomized blocks, 139.
- Balakrishnan, T. R., formation of strata, 186.
- Balanced classifications (equal frequencies), 23; important in Model II, not Model I, 61; robustness, 104-5; *see* Classification, two-way cross-, three-way cross-.

INDEX

- 540
- Balanced incomplete blocks (BIB), 139-46, 151-2, (Exercises 38.10-13, 38.16-18, 38.20) 163-5; unreduced, 142; symmetric, resolvable, affine, dual, 143; complementary, 145; AV, 145-6; with interblock information, 150-1; permutation distributions, 151; preference experiments, 151-2; paired comparisons, 152; confounding, 155.
- Bancroft, T. A., Yates' method of weighted squares of means, 30; orthogonal polynomials in AV, 49; analysis of covariance, 52; pooling in Model II AV, (Example 36.6) 68.
- Banerjee, K. S., decomposition of quadratic forms in normal variables, 4.
- Bargmann, R. E., power of LR test, 282.
- Barnard, G. A., exponential weighting, 501.
- Barnard, M. M., data on Egyptian skulls, (Example 42.3) 277-80.
- Barnett, V. D., canonical analysis in educational research, 305.
- Barra, J. R., review of Latin and orthogonal squares, 137.
- Bartholomew, D. J., ordered alternatives in AV, 49, (Exercise 35.15) 55.
- Bartlett, M. S., transformations, 89, 92, (Exercise 37.6) 115; asymptotic distribution of correlation determinant, (Example 41.4) 249; distribution of covariance, (Exercise 41.13) 262; decomposition, (Exercise 41.16) 262-3; analysis of Egyptian skulls data, (Example 42.3) 277-80; testing latent roots, 292; canonical correlations, 305, 306; factor analysis, 309; discrimination, 326; standard errors of serial correlations, 432, (Exercises 48.2, 48.8-9) 450-1; test for spectral ordinate, 462; smoothing the spectrum, (Example 49.4) 466; multivariate time-series, 496.
- Basu, D., sufficiency in sample survey theory, 170.
- Beale, E. M. L., modified ratio estimator, (Exercise 40.14) 238.
- Behnken, D. W., rotatable designs, 158.
- Belz, M. H., proof of Scheffé's all-contrasts method, (Exercise 35.19) 56.
- Benard, A., ranks test for BIB, 151.
- Bennett, B. M., multivariate test of means, 281, (Exercise 42.12) 283.
- Beveridge, W., wheat-price index, (Table 47.1) 406.
- Bhappkar, V. P., median test, 109.
- Bhuchongkul, S., median estimators in AV, 111.
- Bias, transformation, 95-6.
- Bickel, P. J., median estimators in AV, 110.
- Biggers, J. D., missing observations, 112.
- Binomial, transformations, 95, (Exercises 37.4, 37.6) 114-15; sampling, (Exercises 39.13, 39.23) 207-8; *see also* Negative binomial.
- Birkhoff, G. D., ergodic theorem, 407, 410.
- Bivariate normal distribution, variance-stabilizing transformation of the sample correlation coefficient, (Example 37.3) 93; Wishart distribution of covariances and variances, (Example 41.1) 245; moments of covariance, (Example 41.2) 245; null distribution of latent roots of dispersion matrix, (Example 41.7) 258-9; distribution of covariance, (Exercise 41.13) 262.
- Blackith, R. E., discrimination, 326.
- Blackman, R. B., spectrum analysis, 466, 467, 468, (Exercise 49.5) 469.
- Bliss, C. I., discrimination, 331.
- Blocks, 124; and strata, 182; *see* Experiments, Randomized blocks.
- Bose, R. C., Euler's false conjecture, 136; construction of Latin squares, 137; inequality for resolvable BIB, 143, (Exercise 38.18) 165; construction of BIB, 143; linked paired-comparison designs, 152; PBIB, 153; rotatable designs, 158; Mahalanobis's D^2 statistic, 259; non-central distribution of mean vector, (Exercise 41.1) 260.
- Box, G. E. P., approximate F -test in AV, (Exercise 36.7) 83; distribution of linear function of χ^2 variables, (Exercise 36.14) 84; transformations, 85-8, (Exercises 37.1, 37.9) 114-15; analysis of residuals, 97; robustness of AV, 98, (Exercises 37.10-12) 115-16; evolutionary operation, 158; approximations to moments of LR tests, 268-9; exponential weighting, 501.
- Bozivich, H., pooling in Model II AV, (Example 36.6) 68.
- Bradru, D., non-orthogonal additive multi-way cross-classification, 38.
- Brillinger, D. R., polyspectra, 492.
- Brown, G. W., median tests, 108-9, (Exercises 37.18-20) 117-18.
- Brown, R. G., exponential weighting, 501.
- Brunt, D., rainfall data, (Table 45.2) 343.
- Budne, T. A., random balance experiments, 130.

- Bulmer, M. G., confidence intervals in Model II AV, 71, (Exercise 36.15) 84.
- Burman, J. P., seasonal variation, 400.
- Buys-Ballot table, (Exercises 49.9-10) 470-1.
- Canonical, variables, 285-313; correlations, 299-306, (Exercises 43.5-8, 43.10) 311-13; standard errors, 304-5; *see* Component analysis, Factor analysis, Latent roots.
- Carter, R. L., rotatable designs, 158.
- Chacko, V. J., ordered alternatives in AV, 49.
- Circular, serial correlation, 362; processes, 426; *see* Time-series.
- Classification, *see* Discrimination and classification.
- Classification, hierarchical (nested) in Model I AV, 31-4; two-way, (Example 35.5) 31; three-way, 34, (Exercise 35.3) 53; in Model II, 70-1; balanced two-way (Example 36.7) 70; balanced three-way, (Example 36.8) 70-1.
- Classification, mixed cross- and hierarchical, 34-5.
- Classification, multi-way, in Model I AV, 34-40, (Exercise 35.6) 53; in Model II, 68-70, (Exercise 36.8) 83; permutation tests, 108.
- Classification, one-way, in Model I AV, (Example 35.1) 6, (Exercise 35.1) 52; balanced in Model II, (Examples 36.1-3, 36.5) 58, 61, 64, 68, (Exercises 36.1, 36.3, 36.5, 36.10-11, 36.13) 82-4; Model II unbalanced, 72-3, (Exercises 36.12-13) 83-4.
- Classification, ordered and metrical, 49, (Exercise 35.15) 55.
- Classification, three-way cross-, in Model I AV, 35-9; balanced, (Example 35.6) 38, (Exercise 35.8) 54; disproportional frequencies, 38; balanced, in Model II, 69-70; balanced in mixed model, 77; permutation tests, 108.
- Classification, two-way cross-, in Model I AV, 10-31; proportional frequencies, 18, (Example 35.2) 19; equal frequencies (balanced), (Example 35.3) 23, (Exercise 35.2) 52; disproportional frequencies, 26, (Example 35.4) 27, 28-30, (Exercises 35.4-5, 35.7) 53; empty cells, 30; balanced in Model II, (Examples 36.4, 36.6) 65, 68, (Exercise 36.6) 83; in general model, 75-9; permutation tests, 105-7; median tests, 109, (Exercises 37.18-20) 117-18.
- Classified data, AV for, 11; *see* Classification, one-way, etc.
- Clatworthy, W. H., PBIB, 153.
- Clustering, in sample design, 187-9, 194, (Exercise 39.22) 208; in classification, 336-9.
- Cobb, W., mixed model for AV, 79.
- Cochran, W. G., combination of AV tests, 42; robustness of AV, 98; inference in sample surveys, 119; BIB plans, 143; BIB analyses, 146; lattice designs, 154; confounding, 157; sequences of experiments, 158; sample surveys, 166; formation of strata, 185-6, (Exercise 39.18) 208; systematic sampling, 188; ratio and regression estimators, 223; two-phase sampling, 228; domains of study, 229; random formation of strata, (Exercise 40.6) 236; discrimination, 327, (Example 44.5) 328-9, 329, 331, (Exercises 44.6-9) 340-1.
- Cochrane, D., autocorrelated errors in regression, 497-8.
- Collier, R. O. Jr., robustness of F -test in randomized blocks, 139.
- Combination of AV tests, 40-3.
- Complete randomization, 79.
- Complete sets of orthogonal Latin squares, 136.
- Completeness, in extended exponential family, 67.
- Component analysis, 285-91; latent roots and vectors, geometrical interpretation, 286-9; standardization, 289; testing latent roots, 291-4; and index numbers, 295; meteorological data, (Example 43.4) 295-300; *see* Factor analysis.
- Components, variance, 58; *see* Analysis of variance in Model II; canonical, *see* Component analysis.
- Concomitant variables, 50, 211 f.n.
- Confounding, 155-8; in factorial experiments, 157.
- Contrasts, 46; simultaneous confidence intervals for all, 46-9, (Exercises 35.11-14, 35.16-19) 54-6.
- Cornfield, J., expected MS in AV, 69; models in AV, 75, 77.
- Correlation coefficient, variance-stabilizing transformation in normal samples, (Example 37.3) 93; canonical, 299-306; *see* Autocorrelation, Serial correlation.
- Correlation determinant, moments and asymptotic distribution, (Example 41.4-5) 248-50.

- Correlogram, 362, 404; and spectrum, 410; *see* Time-series.
- Cost function, in survey theory, 182-3; for stratified sample, 183; for multi-stage sample, 201-2; for selection probabilities, 202-4; for two-phase sampling, 226-7; for non-response, (Exercise 40.7) 236.
- Covariance, analysis of, 49-52, 105, 211 f.n.
- Cowden, D. J., moving averages, 374.
- Cox, D. R., sequential AV, 79; transformations, 85-8, (Exercise 37.1) 113; analysis of residuals, 97; experimental inference, 123; expected MS in Latin squares, 138; exponential weighting, 501.
- Cox, G. M., BIB plans, 143; BIB analyses, 146; lattice designs, 154; confounding, 157; sequences of experiments, 158.
- Craddock, J. M., meteorological data component analysis, (Example 43.4) 295-300.
- Cross-classification, *see* Classification, two-way cross-, etc.
- Curtiss, J. H., transformations, 93.
- Dalenius, T., formation of strata, 184-6, (Exercise 39.17) 208.
- Daly, J. F., tests of independence unbiased, 271.
- Daniell, P. J., smoothing the spectrum, (Example 49.5) 466.
- Daniels, H. E., approximations to serial correlation distributions, 446-9; spectrum analysis, 467.
- Darling, D. A., combination of AV tests, 42.
- Darroch, J. N., power of multivariate tests, 271.
- Das Gupta, S., power functions of multivariate tests, 281.
- David, H. A., paired comparisons, 152; polynomial regression designs, 161.
- Davies, O. L., confounding, 157; evolutionary operation, 158.
- Davis, H. T., test in harmonic analysis, 462.
- Decomposition of non-central quadratic forms, 2-5.
- Dempster, A. P., random allocation experiments, 130; multivariate estimation, 264, 291, 306.
- Design, problems, 119; equation, 141; *see* Experiments, Surveys.
- Difference-sign test, 355-7, 360, (Exercises 45.2-3, 45.6) 363-4.
- Discrimination and classification, 314-41; linear discrimination, 314-22; geometrical interpretation, 317, 326; quadratic, 322; testing, 322-3; with cost function, 323-4; k populations, 324-6; qualitative data, 326-7; reserved judgement, 327-8; biased estimation of misclassification errors, 328-329; redundant variables, 329; standard errors of coefficients, 329-30; distribution-free methods, 323, 331-5; differences in dispersion, 335-6; classification, 336-9.
- Dispersion matrix, latent roots of, 255; estimated from residuals, 274, 275-6; *see* Generalized variance (dispersion determinant), Latent roots.
- Distribution-free methods, in AV, 105-11; in testing multivariate location, 282; in discrimination, 323, (Exercises 44.10-11) 341.
- Dixon, W. J., serial correlations, 441-2, (Exercise 48.20) 453.
- Dolby, J. L., transformations, 86.
- Domains of study, 229-34; across strata, 229-32, (Exercises 40.15-16) 238; in multi-stage sampling, 232-4.
- Double sampling, *see* Two-phase sampling.
- Draper, N. R., rotatable designs, 158.
- Duncan, D. B., studentized range tests in AV, 45-6.
- Dunn, O. J., confidence intervals for contrasts in AV, 49, (Exercise 35.14) 55.
- Durbin, J., ranks test for BIB, 151, (Exercise 38.17) 164; estimation of variance in multi-stage sampling, 199, 201, 204, (Exercises 39.26-7) 208-9; selection with unequal probabilities, (Exercises 39.2, 39.4) 205; reduction of bias in ratio estimator, 216, 217; asymptotic linearity of ratio and regression estimators, 223; domains of study, 229, (Exercises 40.15-16) 238; non-response in surveys, (Exercise 40.8) 236; spectrum and seasonal variation, 467; moving average and autoregressive schemes, 481-5, (Exercise 50.6) 503; regression with autocorrelated errors, 497-8; mixed autoregressive-regressive systems, 499; Yule-Walker equations, (Exercise 50.5) 502.
- Eisenhart, C., AV models, 57.
- Eisenpress, H., seasonal variation, 400.
- Ekman, G., formation of strata, 185.
- Empty cells in cross-classified data, 30.
- Equivalent samples, 170.
- Ergodic, 407-8, 410; *see* Time-series.

INDEX

- Errors, unit, interactive and technical, 80.
 Euler, Latin squares, 134; false conjecture, 136.
 Evolutionary operation, 158.
 Experiments, design of, 119-65; compared with surveys, 119, 182; principles, randomization, 120-5; block experiments, 124-30; incidence matrix, 125; linear model, 125-9; AV, 128-9; design, 129-30; with two nuisance factors, 132-4; interblock information, 146-51; preference, 151-2; factorial, 154-5; confounding, 155-8; sequences of, 158; regression, 158-61; *see* Balanced incomplete blocks, Latin squares, Randomized blocks
- Factor analysis, 306-11 (Exercises 43.11-13) 313; indeterminacy resolved, 307; ML solution, 308-9; tests for factors, 309-10; discussion, 310-11; *see* Component analysis.
- Factorial experiments, 154-5; confounding, 157; fractionally replicated, 157.
- Fellegi, I., sampling with unequal probabilities, 171, 174.
- Filter, 423-4; *see* Time-series.
- Finch, P., autoregressive and moving average schemes, 484.
- Finite populations, 166; *see* Surveys.
- Finney, D. J., combination of AV tests, 42; probit and logit transformations, 95.
- Fisher, R. A. (Sir Ronald), originator of AV, 2; LSD test in AV, 44; transformation of r , 92-3; test of symmetry, 106; advocacy of randomization, 120; experimental inference, 123; BIB inequality, 142; confounding, 157; latent roots, 258; discrimination, (Example 44.1) 317-18, 339; test in harmonic analysis, 462.
- Fortier, J. J., cluster analysis, 337.
- Foster, F. G., distribution of latent roots, 259, (Example 42.4) 280-1; records test, 360, (Exercises 45.8-9) 364-5.
- Freeman, G. H. and Jeffers, J. N. R., non-orthogonal three-way cross-classification, 38.
- Freeman, M. F., transformations, 90, (Exercise 37.4) 114.
- Friedman, M., AV using ranks, (Exercise 37.14) 116.
- Gabriel, K. R., AV of cell means, 38; simultaneous and step-by-step tests in AV, 45, 48-9, (Exercises 35.12, 35.17-18) 55-6.
- Gamma distribution, logarithmic transformation, (Example 37.2) 91-2.
- Gardiner, D. A., rotatable designs, 158.
- Gassner, B. J., BIB designs, (Exercise 38.20) 165.
- Gautschi, W., completeness, 62, 67.
- Gaylor, D. W., polynomial regression designs, 161.
- Geary, R. C., distribution of ratio, 446.
- Geisser, S., multivariate normal theory, (Exercise 41.5) 260-1.
- General mean, 12.
- Generalized variance (dispersion determinant), distribution and moments, (Example 41.3) 246-8, (Example 41.5) 249-50, (Exercises 41.8-9) 261; estimation of, 264.
- Ghosh, B. K., sequential AV, 79.
- Ghosh, M. N., Tukey's test for additivity, 25.
- Ghosh, S. P., formation of strata, 186; two-phase sampling for clustering, 228.
- Girshick, M. A., latent roots, 258, 293-4, (Exercise 43.9) 312; canonical correlation, (Exercise 43.8) 312, (Exercise 43.10) 313.
- Gleissberg, W., test of randomness, 354.
- Gleser, L. J., sphericity test, 272.
- Godambe, V. P., linear estimation in sample surveys, 174.
- Good, I. J., circulant matrices, 394, (Exercise 46.13) 402.
- Goodman, L. A., unbiased ratio estimators, 214, 215, 217.
- Graeco-Latin squares, 135-6, (Exercise 38.6) 162; factorial experiments in, 155.
- Grandage, A. H. E., rotatable designs, 158.
- Granger, C., spectrum analysis, 468, 469, 492.
- Graybill, F. A., decomposition of quadratic forms in normal variables, 4; Model II AV, 59, 62, 63; inter-block information, 150.
- Grenander, U., spectrum analysis, 468; regression with autocorrelated errors, 498.
- Grundy, P. M., sampling with unequal probabilities, 173, (Exercise 39.11) 206-7.
- Guérin, R., review of BIB and PBIB designs, 142-3, 153.
- Guest, P. G., polynomial regression designs, 161.
- Haavelmo, T., systems of equations, 499.
- Hader, R. J., rotatable designs, 158.
- Hájek, J., limiting normality, 169; rejective sampling, 174.
- Hanani, H., BIB inequality, 142.

- Hannan, E. J., spectrum and seasonal variation, 467; regression with autocorrelated errors, 497.
- Hansen, M. H., sample surveys, 166; choice of selection probabilities, 202; non-response in sample surveys, (Exercise 40.7) 236.
- Harman, H. H., factor analysis, 310.
- Harmonic analysis, *see* Spectrum.
- Harter, H. L., multiple comparisons methods, 46.
- Hartley, H. O., combination of AV tests, 40-3; Newman-Keuls test, 46, (Exercise 35.10) 54; pooling in Model II AV, (Example 36.6) 68; unbiased ratio estimators, 212, 214, 215, 217; domains of study, 229; random formation of strata, (Exercise 40.6) 236.
- Hatanaka, M., spectrum analysis, 468, 492.
- Healy, M. J. R., transformation tables, 87; data on male premolars, (Example 42.4) 280-1.
- Henderson, C. R., estimation in unbalanced Model II AV, 73.
- Herbach, L. H., testing hypotheses in Model II AV, 68, (Exercises 36.5-6) 82-3.
- Hess, I., formation of strata, 186; ratio estimation, 223.
- Hext, G., spectrum analysis, 467.
- Hierarchical classification, *see* Classification, hierarchical.
- Higham, J. A., trend fitting, (Exercise 45.6) 400-1.
- Hills, M., discrimination, (Example 44.5) 329.
- Hodges, J. L., median estimators in AV, 110; formation of strata, 185.
- Hoel, P. G., regression designs, 161; distribution of dispersion determinant, (Exercise 41.8) 261.
- Hogg, R. V., nested hypotheses, (Exercise 37.2) 114; testing degree of polynomial regression, (Exercise 37.3) 114; power of tests, 281.
- Holt, C. C., exponential weighting, 501.
- Homogeneity, LR tests of, 87, (Exercises 37.1-3) 113-14, 264-9, (Example 42.1) 270, (Example 42.2) 272-3, (Exercises 42.1-3) 282.
- Hopkins, C. E., discrimination, 327, (Example 44.5) 328-9.
- Horsnell, G., robustness of AV, 98.
- Horvitz, D. G., sampling with unequal probabilities, 173.
- Hotelling, H., T^2 as AV test, 79; variance-stabilization, 92; T^2 distribution, 250, 253, (Exercise 41.12) 262; geometrical interpretation, 251-2; T^2 and R^2 , 252; T^2 and F , 252; one-sample T^2 test, 252; two-sample T^2 test, 253; two-sample T^2 and D^2 , 260; non-central T^2 , 259, 281; T_0^2 test for several samples, 281-2; robustness and power of T^2 , 281-2, (Exercises 42.11, 42.13-14) 283; T^2 as LR test, (Exercise 42.10) 283; T^2 in generalization of Scheffé's test, (Exercise 42.12) 283; canonical correlations, 299, (Example 43.5) 303-4, 305, (Exercises 43.5-7) 311-12.
- Howe, W. G., factor analysis, 309.
- Høyland, A., median estimators in AV, 110.
- Hsu, P. L., latent roots, 258; multivariate Beta distribution, 260.
- Hultquist, R. A., Model II AV, 59, 62, 63.
- Hunter, J. S., response surfaces, 158.
- Hurwitz, W. N., sample surveys, 166; choice of selection probabilities, 202; non-response in sample surveys, (Exercise 40.7) 236.
- Imhof, J. P., mixed model in AV, 77.
- Incidence matrix, of an experiment, 125.
- Independence, LR test of, (Exercises 41.10-11) 261-2, 270-1, 281, (Exercise 42.4) 282; equivalent to testing equality of latent roots, 291.
- Index number, from component analysis, 295.
- Intensity, 411; *see* Spectrum, Time-series.
- Interactions in AV, 13, 36; Tukey's test for, (Example 35.3) 23; zero for any weights if for one set, 26; independent and tied, 75-6; unit-treatment, 80.
- Interactive errors, 80.
- Inter-block information, 146-51, (Exercises 38.14-16) 164.
- Inverse sampling, (Exercise 40.3) 235.
- Ito, K., robustness of T^2 test, 281; power of tests for mean-vectors, 281-2.
- James, A. T., latent roots, 260.
- Jeffers, J. N. R., *see* Freeman, G. H.
- Jenkins, G. M., joint distribution of serial correlations, 437, 449, (Exercises 48.18-48.19) 453; non-null distribution of serial correlations in Markoff case, 444, 445, (Exercise 48.17) 452; spectrum analysis, 466; exponential weighting, 501.

- John, S., discrimination, (Example 44.5) 329, (Exercise 44.2) 339.
- Johnson, A. H. L., square root transformations, (Exercise 37.15) 117.
- Johnson, N. L., sequential AV, 79; quota sampling, (Exercise 40.11) 237-8; variate-difference method, 391.
- Kelley, T. L., psychological data for canonical analysis, (Example 43.5) 303.
- Kempthorne, O., models in AV, 75; complete randomization, 81; expected mean squares in randomized blocks and Latin squares, 138; BIB analyses, 146; PBIB analysis, 153; lattice designs, 154; confounding, 157; sequences of experiments, 158.
- Kendall, D. G., logarithmic transformation, 92.
- Kendall, M. G., AV using ranks, (Exercise 37.14) 116; n -dimensional geometry, 243 f.n.; computation in component analysis, 289; ranking for principal components, 295; factor analysis, 310; classification, (Example 44.7) 338-9; central limit for moving average weights, 370 f.n.; bias in serial correlations, 435, (Exercises 48.4, 48.11) 450-1; distribution of serial correlation in Markoff case, 444.
- Keuls, M., studentized range test in AV, 45-6, (Exercise 35.10) 54.
- Khamis, S. H., sample survey theory, 170, (Exercise 39.1) 204.
- Khintchin, A., ergodic theorem, 407, 410.
- Kiefer, J., optimal experiments, 130; regression designs, 158.
- Kish, L., estimation of variance, (Exercise 39.16) 207; ratio estimation, 223.
- Koop, J. C., linear estimation in sample surveys, 174.
- Koopmans, T. C., serial correlations, 442.
- Kruskal, J. B., monotone transformations, 88.
- Kshirsagar, A. M., multivariate Beta distribution, 260; Bartlett decomposition, (Exercises 41.16-17) 262-3.
- Lahiri, D. B., selection with unequal probabilities, (Exercise 39.11) 206-7; removal of bias in ratio estimators, 223.
- Latent roots of a dispersion matrix, null distribution, 255-8, 259, 260, (Exercise 41.14) 262; (Example 42.4) 280-1; testing equality equivalent to testing independence, 291, (Example 43.3) 292; testing zero value, 292; testing equality of small roots, 292-3; large-sample results, 293-4; in discrimination, 326; *see* Canonical correlations, Component analysis, Factor analysis.
- Latin squares, 134-7, (Exercises 38.5-7) 162; robustness of normal theory, 138-9; factorial experiments in, 155.
- Lattice designs, 153-4.
- Lawley, D. N., distribution of LR statistic, 269; T_0^2 test, 281; component analysis, (Example 43.2) 290-1, 293; canonical correlations, 305, 306; factor analysis, 308, 310, (Exercise 43.11) 313.
- Least squares, in sampling without replacement, 167-8; in discrimination, 323, (Exercises 44.10-11) 341; for moving averages, 366-7, 374-5; in autoregressive series, 476, 499; *see* Linear model.
- Ledermann, W., component analysis, 285.
- Lehmann, E. L., median estimators in AV, 110; multivariate tests, 281.
- Leipnik, R. B., serial correlation, 444, 445.
- Levene, H., tests of randomness, 354, 355, (Exercises 45.6-7) 364.
- Levine, A., polynomial regression designs, 161.
- Lewis, T., canonical analysis in educational research, 305.
- Likelihood Ratio (LR) tests, in Model II AV, (Exercises 36.5-6) 82-3; for "nested" hypotheses, 87, (Exercises 37.1-3) 113-14; in multivariate analysis, 265-84; *see* Homogeneity, Independence, Regression, Sphericity.
- Linear model, AV in (Model I), 1-56; decomposition of non-central quadratic forms, 2-5; removal of singularity, 12; choice of weights, 25-6; general disproportional frequencies case, 38; combination of tests, 40-3; multiple comparisons, 43-9; analysis of covariance, 49-52; extension of model to further parameters, 51-2; transformations to, 85-8; missing observations, 111-13; for block experiments, 125-9; in sampling without replacement, 167-8; multivariate, 273-6, (Example 42.3) 277-80; *see* Analysis of Variance, Classification, one-way, etc.
- Linhart, H., discrimination, 327.
- Lipton, S., data on male premolars, (Example 42.4) 280-1.
- Logarithmic transformations, (Examples 37.2, 37.4) 91, 93; 95-6.

INDEX

546

- Logit transformations, 94.
- Lomnicki, Z. A., test for normality of a stationary process, (Exercise 49.8) 470.
- Low, L. Y., estimation in unbalanced Model II AV, 73.
- LSD test in AV, 93-4.
- Lubischew, A. A., discrimination, 336.
- Madow, W. G., sample surveys, 166; limiting normality, 169; serial correlation, 443.
- Mahalanobis, P. C., D^2 statistic and generalized distance, 259-60.
- Main effects in AV, 12, 36.
- Manley, G., meteorological data, (Example 43.4) 295.
- Mann, H. B., complete sets of Latin Squares, 137; construction of BIB, 143; confounding, 157; difference-sign test, 357, (Exercise 45.3) 363; rank correlation test, 358; LS in autoregressive series, 476.
- Markoff series, autocorrelations, (Example 47.2) 405; backwards, (Example 47.3) 406; correlogram and spectrum, (Example 47.7) 418-19; partial autocorrelations, 424-5; cumulants and normality, (Exercise 47.2) 426; (Exercise 47.3) 427; grouping, (Exercise 47.15) 428; standard error of serial correlations, (Example 48.3) 432; covariance of serial correlations, (Example 48.4) 433; bias in serial correlation, (Example 48.7) 435, (Exercises 48.4-5; 48.8, 48.11) 450-1; to higher order, 435; non-null distribution of serial correlations, 443-4, 447-9, (Exercises 48.13, 48.17) 451-2; effect of starting-point, 472-4; multivariate, 496; in forecasting, 501.
- Marriott, F. H. C., bias in serial correlations, 435.
- Marsaglia, G., decomposition of quadratic forms in normal variables, 4.
- Mauchly, J. W., sphericity test, 271.
- Mauldon, J. G., multivariate estimation paradoxes, 281.
- Maxwell, A. E., component analysis, (Example 43.2) 290-1; factor analysis, 308, 310, (Exercise 43.11) 313.
- Mean, moments of, 168-9.
- Mean squares (MS), expected values of in Model II AV, 63, 69.
- Median tests in AV, 108-11, (Exercises 37.18-20) 117-18.
- Mickey, M. R., unbiased regression-type estimators, 219.
- Mikhail, N. N., power of tests for mean-vectors, 281-2.
- Mill, J. S., on experiments, 120.
- Mixed models, 77-9; for recovery of inter-block information, 146-51.
- Models I, II; *see* Analysis of variance.
- Mood, A. M., median tests, 109-11, (Exercises 37.18-20) 117-18; latent roots, 258.
- Moore, G. H., tests of randomness, 354, 356.
- Moran, P. A. P., Slutsky sinusoidal limit, 415; moments of serial correlations, 435-7, (Exercises 48.5-6, 48.12) 450-1.
- Moving average, 367-402; as LS polynomial, 366-7; formulae to degree 5, 368-9; formulae in terms of differences, 370; Spencer's 15- and 21-point formulae, (Examples 46.3-4) 372; end-effects, 373-4; using orthogonal polynomials, 374-5; *see* Moving average series, Seasonal variation, Trend.
- Moving average series, 412-16; and autoregressive series, 417, 474-6; estimates and tests of fit, 481-4, (Exercise 50.6) 503; as errors in autoregressive series, 484-6; exponentially weighted, 501; *see* Autoregressive series, Time-series.
- Mudholkar, G. S., power functions of multivariate tests, 281.
- Muller, E.-R., BIB designs, 143.
- Multinormal distribution, multivariate normal distribution; *see* Multivariate analysis.
- Multi-phase sampling, 228.
- Multiple comparisons in AV, 43-9.
- Multi-stage sampling, 189-204, (Exercises 39.24, 39.28-9) 208-9; estimator, 191; with equal probabilities, 191-4; with unequal probabilities, 195-7; estimation of variance, 197-201, 223-4; cost function and minimum variance, 201-2; choice of probabilities, 202-4; efficiency, 204; stratification, 204; ratio and regression estimators, 223-4; domains of study, 232-4.
- Multivariate analysis, 239-341; in time-series, 486-96; *see* Canonical variables, Component analysis, Correlation determinant, Discrimination and classification, Dispersion matrix, Factor analysis, Generalized variance, Homogeneity, Hotelling T^2 , Independence, Latent roots, Regression, Sphericity, Wishart distribution.
- Murteira, B., variate-difference method, (Exercises 48.15-16) 452.

- Murthy, M. N., sample survey theory, 166; sufficiency in surveys, 176, (Exercises 39.30-1) 209-10; unbiased ratio estimators, 223, (Exercise 40.1) 234.
- Murty, V. N., inequality for BIB, (Exercise 38.13) 163.
- Nair, K. R., PBIB, 153.
- Nanjamma, N. S., unbiased ratio estimators, 223, (Exercise 40.1) 234.
- Narain, R. D., tests of independence unbiased, 271.
- Negative binomial distribution, angular transformation, 95, (Exercise 37.5) 114.
- Nerlove, M., spectrum analysis, 467.
- Nested classification, 31 f.n.; *see* Classification, hierarchical.
- Nested hypotheses, 87, (Exercises 37.1-3) 113-14.
- Newman, D., studentized range test in AV, 45-6, (Exercise 35.10) 54.
- Neyman, J., transformation bias, 95; stratified sampling, 180; two-phase sampling, 224.
- Nieto de Pascual, J., unbiased ratio estimators, 213, 215, 217, 223.
- Noether, G. E., ranks test for BIB, 151; rank serial correlation test, 360.
- Non-central quadratic forms, decomposition of, 2-5.
- Normal distribution, logarithmic transformation of sample variance, (Examples 37.2, 37.4) 91, 93; square root transformation of sample variance, (Example 37.5) 94; *see also* Bivariate normal, Multivariate analysis.
- Normal scores, transformation to, 94; AV using, 105, 107-8.
- Normalizing transformations, 93-4.
- Norton, H. W., review of Latin squares, 137.
- Nuisance factors, 124; two, 132; three or more, 135-7.
- Ogawa, J., robustness of F -test in randomized blocks and BIB, 139, 151.
- Olkin, I., multivariate ratio estimators, 216, 223.
- One-way classification, *see* Classification, one-way.
- Orcutt, G. H., autocorrelated errors in regression, 497-8.
- Orthogonal squares, 136, (Exercises 38.6-7) 162; factorial experiments in, 155.
- Paired comparisons, 152.
- Parker, R., Euler's false conjecture, 136.
- Partially balanced incomplete blocks (PBIB), 152-3.
- Parzen, E., spectrum analysis, 466, 467, (Exercises 49.6-7) 470.
- Pathak, P. K., sufficiency in sample survey theory, 170-1, (Exercise 39.30) 209, 223, (Exercise 40.3) 235.
- Patterson, H. D., sampling on successive occasions, (Exercises 40.9-10) 236-7.
- Pearce, S. C., review of non-orthogonal AV, 38.
- Pearson, E. S., homogeneity tests, (Example 42.2) 272.
- Periodogram, 411-12; *see* Spectrum, Time-series.
- Permutation tests, in AV, 105-8, 138-9, 151.
- Phases, in time-series, 353-5, (Exercise 45.1) 363; in harmonic analysis, 454; *see* Two-phase sampling, Multi-phase sampling.
- Phillips, A. W., multivariate time-series, 489.
- Pillai, K. C. S., distribution of latent roots, 259, 260.
- Pitman, E. J. G., permutation test in AV, 106-7, (Exercise 37.13) 116.
- Plackett, R. L., models in AV, 75, 81; duplicated observations in AV, 113.
- Poisson, distribution, square root transformations, (Example 37.1) 89-90, (Exercises 37.15-17) 117; sampling, (Exercises 39.13, 39.23) 207-8.
- Polynomial, testing degree in regression, (Exercise 37.3) 114; regression designs, 158-61.
- Pooling procedures, in AV, (Example 36.6) 68; in regression, (Exercise 37.3) 114.
- Pope, J. A., bias in serial correlations, 435.
- Posten, H. O., power of LR test, 282.
- Preference experiments, 151-2.
- Principal components, 287; *see* Component analysis.
- Probabilities proportional to size (p.p.s.), 195-7, 204; *see* Unequal probabilities, Surveys.
- Probit transformations, 94.
- Product estimator, (Exercise 40.2) 234-5.
- Puri, M. L., median estimators in AV, 111.
- Quadratic forms, decomposition of, 2-5.
- Quasi-random sampling, 188.
- Quenouille, M. H., sequences of experiments, 158; method of bias-reduction, 216, 264, 306, 435; variate-difference method, 393, (Exercises 46.7-11) 401; trend-fitting,

- 393-4, 396, (Exercise 46.12) 402; large-sample theory of serial correlations for autoregressive series, 433, (Exercise 48.14) 452; non-null distribution of serial correlation in Markoff case, 444, transformed, (Exercise 48.13) 451-2; joint distribution of serial correlations, 449; robustness of serial correlation theory, 449; unequal time-intervals in time-series, 469; partial autocorrelations and test of fit in autoregressive series, 478; multivariate time-series, 486-9, 492; series with common errors, (Exercise 50.8) 503.
- Quota sampling, (Exercise 40.11) 237-8.
- Raj, D., sufficiency in surveys, 170, (Exercise 39.1) 204; unequal probabilities, 176, (Exercises 39.9-10) 206; two-phase sampling for probabilities, 228.
- Rajalakshman, D. V., multivariate time-series, 496.
- Randomization, complete, 79; in experiments, 120-5.
- Randomized blocks, 79-80, 130-2, (Exercises 38.3-4) 162; robustness of normal theory, 138-9; factorial experiments in, 155.
- Randomness, tests of, 360; *see* Difference-sign, Phases, Rank correlation, Records, Serial correlation, Turning-points.
- Range tests in AV, 44-6.
- Rank, transformations, 94; AV using, 105, 107-9, (Exercises 37.13-14) 116; test in BIB, 151; correlation tests in time-series, 357-60; serial correlation test, 360, (Exercise 45.5) 363.
- Rao, C. R., BIB designs, 143-4; BIB analyses, 146; PBIB analysis, 153; discrimination, 324, (Example 44.4) 325-6.
- Rao, J. N. K., unequal probabilities, 174; reduction of bias in ratio estimator, 216; random formation of strata, (Exercise 40.6) 236.
- Ratio estimators, biasedness, 211-12, 216-18, 222-3; consistency, 212; modified, 212-13, (Exercises 40.1-2) 234-5, (Exercises 40.13-40.14) 238; variance comparisons, 213-18; in stratified and multi-stage sampling, 223-4; asymptotically linear, 223-4; in two-phase sampling, 227-8.
- Realization, 404.
- Recognizable individuals, in sample survey theory, 166, 170-1, 174-5.
- Records test, 360, (Exercises 45.8-9) 364-5.
- Recovery of inter-block information, *see* Inter-block information.
- Rees, D. H., non-orthogonal additive multi-way cross-classification, 38; distribution of latent roots, 259, (Example 42.4) 280-1.
- Regression, testing degree of polynomial, (Exercise 37.3) 114; transformations, (Exercise 37.9) 115; designs, 158-61; in multivariate analysis, 273-6, (Example 42.3) 277-80, (Exercises 42.15-16) 283-4; with autocorrelated errors, 497-8; *see* Autoregressive series.
- Regression estimators, 218-19; unbiased, 219-22; in stratified and multi-stage sampling, 223-4; asymptotically linear, 223-4; in two-phase sampling, 227-8.
- Rejective sampling, 174.
- Replacement, sampling with and without, 166; *see* Surveys.
- Residuals, analysis of, 96-7; dispersion matrix of, 274, 275-6.
- Response surfaces, 158.
- Rhodes, E. C., trend-fitting, 393.
- Robson, D. S., ratio estimators, 214, 223; product estimator, (Exercise 40.2) 235-6.
- Robustness, of AV procedures, 97-108.
- Romanovsky, V., Slutsky sinusoidal limit, 415.
- Rosenblatt, M., spectrum analysis, 468; regression with autocorrelated errors, 498.
- Ross, A., allocation in stratified sampling, (Exercise 39.20) 208; unbiased ratio estimators, 212.
- Rotatable designs, 158.
- Roy, J., inter-block information, 151.
- Roy, S. N., mixed model in AV, 79; distribution of latent roots, 258, 259; Mahalanobis's D^2 statistics, 259.
- Rubin, H., serial correlations, 442.
- Sampford, M. R., inverse sampling, (Exercise 40.3) 235.
- Sample surveys, *see* Surveys.
- Satterthwaite, F. E., approximate F -test in AV, (Exercise 36.7) 83; random balance experiments, 130.
- Schatzoff, M., comparison of tests for mean-vectors, 282.
- Scheffé, H., Tukey's test for additivity, 25; interactions zero for any weights if for one set, 26; analysis of cross-classified data with empty cells, 30; three-way hierarchical classification, 34; multiple comparisons, 46; simultaneous confidence

- intervals for all contrasts, 48-9, (Exercises 36.11-13) 54-5; analysis of covariance, 52; expected MS in AV, 69; Model II three-way hierarchical classification, 71; confidence intervals in Model II AV, 71, (Exercise 36.15) 84; models for AV, 75; mixed model, 75, 77, 79; robustness of AV, 98; AV of cell means, (Exercise 37.7) 115; interaction in Latin squares, 135; robustness in randomized blocks and Latin squares, 138-9; problem of two means, 281.
- Schmetterer, L., transformation bias, 95.
- Schull, W. J., robustness of T^2 test, 281.
- Scott, E. L., transformation bias, 95.
- Searle, S. R., estimation in unbalanced Model II AV, 73.
- Seasonal variation, 349-50, 396-400, 403; and spectrum, 467-8; *see* Moving average, Trend.
- Seber, G. A. F., orthogonality in AV, 37; power of multivariate tests, 281.
- Self-weighting designs, 195, 202.
- Sen, A. R., selection schemes with unequal probabilities, (Exercises 39.5-6) 205-6.
- Sequential analysis of variance, 79.
- Serial correlation, using ranks, 360, (Exercise 45.5) 363; generally, 361-2, (Exercises 45.10-11) 365; and variances of differences, 391-2; and variate-difference method, 393, (Exercises 46.10-11) 401; large-sample theory, 431-3, (Exercises 48.3, 48.9-10) 450-1; bias, 433-5; exact moments, 435-7, (Exercises 48.5-6, 48.12, 48.18-19) 450-3; distribution in normal case, 437-49, (Exercise 48.20) 453; transformations, 445, (Exercise 48.13) 451-2; *see* Autocorrelation.
- Seshadri, V., inter-block information, 150.
- Sethi, V. K., formation of strata, 186; unbiased ratio estimators, 223.
- Shah, K. R., inter-block information, 151.
- Sharma, D., Tukey's test for additivity, 25.
- Shiskin, J., seasonal variation, 400.
- Shrikande, S. S., Euler's false conjecture, 136; PBIB, 153.
- Silvey, S. D., power of tests, 281.
- Simaika, J. B., power of T^2 test, 281.
- Simultaneous test procedures, 44-9, (Exercises 35.11-14, 35.16-17, 35.19) 54-6.
- Siotani, M., confidence intervals for contrasts in AV, 49.
- Sitgreaves, R., discrimination, (Exercise 44.2) 339.
- Slater, P., discrimination data on neurotics, (Example 44.4) 325-6.
- Slutzky-Yule effect, 378; sinusoidal limit theorem, (Example 47.6) 414-15.
- Smith, B. Babington, AV using ranks, (Exercise 37.14) 116.
- Smith, C. A. B., quadratic discrimination, 322.
- Smith, K., polynomial regression designs, 161, (Exercise 38.19) 165.
- Snedecor, G. W., Yates' method of weighted squares of means, 30.
- Solomon, H., cluster analysis, 337.
- Spectrum, spectral density, spectral function, 410-11; as autocorrelation g.f., 410; of Markoff and Yule series, (Examples 47.7-8) 418-20; for continuous series, 422; effect of filtering, 423-4; analysis, 454-71; harmonic analysis, 454-5; Nyquist frequency, aliases, 455-7; effect of a harmonic component, 458-60; effect of other periodicities, trend, 460-1; test for the spectral ordinate, 461-2; smoothing, 463-4; calculation of, 464-6; estimation of density, 466-7; and seasonal variation, 467-8; unequal time-intervals, 468-9; cross-spectra, 491-6; coherence, 491; polyspectra, 492; *see* Time-series.
- Spencer's 15- and 21-point formulae, (Examples 46.3-4) 372.
- Sphericity test, 271-2, (Example 43.3) 292.
- Split-plot designs, 157.
- Sprott, D. A., BIB designs, 143-4; recovery of inter-block information, (Exercise 38.16) 164.
- Square root transformations, (Example 37.1) 89-90, 95, (Exercises 37.15-17) 117.
- Srivastava, S. R., pooling procedures in Model II AV, (Example 36.6) 68.
- Stabilization of variance, 88-92.
- Stages, *see* Multi-stage sampling.
- Stationary time-series, 404; *see* Time-series.
- Step-by-step AV test procedures, 42-6, (Exercise 35.18) 56.
- Stevens, W. L., non-orthogonal three-way cross-classification, 38.
- Stratified sampling, 177-87, (Exercises 39.13-39.15, 39.17-21) 207-8; motivation for, 177-9; choice of sample sizes, 180-2; strata and blocks, 182; MV allocation for fixed cost, 183; formation of strata,

- 183-6, (Exercises 40.4-6) 235-6; estimation of effect, 186-7; and clustering, 188-9; in multi-stage sampling, 204; ratio and regression estimators, 223-4; with two phases, 224-6; domains of study, 232-4, (Exercises 40.15-16) 238; quota sampling, (Exercise 40.11) 237-8.
- Stuart, A., random formation of strata, (Exercises 40.4-6) 236; difference-sign test, 357, 360; records test, 360, (Exercises 45.8-9) 364-5; turning-points test, 360, (Exercise 45.4) 363; rank correlation tests, 360; rank serial correlation tests, 360, (Exercise 45.5) 363.
- "Student," (W. S. Gosset), LSD test in AV, 43.
- Studentized range tests in AV, 44-6.
- Successive occasions, sampling on, (Exercises 40.9-10) 236-7.
- Sufficiency, in Model II AV, 62, 73, (Exercise 36.12) 83; in sample survey theory, 170-1, 176, (Exercise 39.1) 204, (Exercises 39.30-1) 209-10.
- Supplementary information, 211-38; *see* Ratio estimators, Regression estimators, Two-phase sampling.
- Surveys, compared with experiments, 119, 182; theory, 166-238; random sampling without replacement, 167-8; moments of sample mean, 168-9; sufficiency, 170-1; *see* Domains of study, Multi-stage sampling, Ratio estimators, Regression estimators, Stratified sampling, Two-phase sampling, Unequal probabilities.
- Sweeny, H. C., polynomial regression designs, 161.
- Systematic sampling, 187-8.
- Tamura, R., multivariate distribution-free location tests, 282.
- Taylor, L. R., transformation tables, 87.
- Technical errors, 80.
- Thompson, D. J., sampling with unequal probabilities, 173.
- Thompson, W. A., Jr., negative estimates of variance in Model II AV, 71.
- Tidwell, P. W., transformations, 86, (Exercise 37.9) 115.
- Tied interactions, 75-6.
- Time-series, 342-503; general, 342-8; components of, 349-50, 366; tests of randomness, 350-61; moving averages, 367-402; stationary, 403-4; ergodic, 407-8, 410; intensity, 411; periodogram, 411-12; moving average series, 412-16; autoregressive series, 416-21; continuous series, 421-3; filters and transfer functions, 423-4; infinite and circular processes, 425-6; serial correlations, 360-2, 431-53; spectrum analysis, 410-11, 454-71; estimation and testing in autoregressive and moving average series, 472-86, 497-500; multivariate, 486-96; systems of equations, 496-7; forecasting, 500-2; *see* Autocorrelation, Autoregressive, Markoff, Moving average series, Randomness, tests of, Serial correlation, Spectrum, Trend, Variate-difference, Yule series.
- Tin, M., ratio estimators, 217-18, (Exercises 40.13-14) 238.
- Tintner, G., variate-difference method, 390, 391.
- Tocher, K. D., missing observations, 112; other spoilt experiments, 113; block experiments, 124, 140, (Exercises 38.2-3, 38.8-9) 162-3; inter-block information, 150, (Exercise 38.15) 164.
- Transfer function, 423-4; *see* Time-series.
- Transformations, to the normal linear model, 85-8; purposes of, 87-8; monotone, 88; variance-stabilizing, 88-92; normalizing, 93-4; to additivity, 94-5; removal of bias, 95-6; analysis of residuals, 96-7; *see also* Angular, Logarithmic, and Square Root transformations.
- Treatments, 124; AV for, 155.
- Trend, 349-50, 366; tests against, 355, 360; effect of elimination by moving averages (Slutzky-Yule effect) 375-84, 393-6, (Exercise 45.12) 402; *see* Moving average.
- Tryon, R. C., cluster analysis, 337.
- Tschuprow, A. A., stratified sampling, 180.
- Tukey, J. W., test for additivity, 25; multiple comparisons, 43-9; studentized range tests, 44, 45; simultaneous confidence intervals for all differences, contrasts, combinations, 46-7, (Exercise 35.13) 55; expected MS in AV, 69; estimation in unbalanced Model II AV, 73; models in AV, 75, 77; moments of variance estimators in AV, (Exercise 36.10) 83; transformations, 90, (Exercise 37.4) 114; analysis of residuals, 96; spectrum analysis, 466, 467, 468, (Exercise 49.5) 469.
- Turning-points test, 351-2, 353, 354, (Exercise 45.4) 363.

- Two-phase sampling, 224-8; for stratification, 224-6; cost function, 226-7; for ratio estimation, 227-8; for regression estimation, 228.
- Unequal probabilities, in sampling without replacement, 171-6, (Exercises 39.2-11) 205-7, (Exercises 39.30-1) 209-10; linear estimation, 173-5; with replacement, 176-177; and stratification, 177-9; and clustering, 187-9; and multi-stage sampling, 189-190, 195-204; p.p.s. sampling, 195-7; estimation of sampling variance, 197-201; chosen to minimize variance, 202-4; chosen to remove bias, 223-4; two-phase sampling to determine, 228; *see* Multi-stage sampling, Stratified sampling, Surveys.
- Uniform sampling fraction (USF), 180.
- Unit errors, 80.
- Van Elteren, P., ranks test for BIB, 151.
- Variance, *see* Analysis of variance.
- Variance-stabilizing transformations, 88-92.
- Variate-difference method, 384-93, (Exercises 46.7-11) 401, (Exercises 48.15-16) 452.
- Verhagen, A. M. W., proof of Scheffé's all-contrasts method, (Exercise 35.19) 56.
- Vithayasai, C., ratio estimators, 223.
- Vos, J. W. E., sampling in time and space, (Exercise 40.10) 237.
- Wagle, B., latent roots distributions, 259.
- Wald, A., discrimination, 339, (Exercise 44.2) 339; rank serial correlation test, 360; LS in autoregressive series, 476.
- Walker, A. M., autoregressive and moving average schemes, 484.
- Walker, G., equations for autoregressive series, 417; test in harmonic analysis, 461.
- Wallis, W. A., tests of randomness, 354, 356.
- Ward, D. H., exponential weighting, 501.
- Watson, G. S., robustness of AV, 98, (Exercises 37.10-12) 115-16; joint distribution of serial correlations, 449; regression with autocorrelated errors, 497.
- Weeks, D. L., inter-block information, 150.
- Weights, choice of, in AV for linear model, 25-6.
- Welch, B. L., robustness of AV, 103, 139.
- White, J. S., bias in serial correlations, 435; moments of serial correlation in Markoff case, 445.
- Whittaker, E. T., periodogram, (Exercise 49.10) 470-1.
- Whittle, P., autoregressive series, (Exercise 47.16) 428; autocorrelation matrix, (Exercise 47.18) 429; estimating spectral density, 467; moving-average series estimator, (Example 50.1) 475-6; estimation and testing in time-series, 486.
- Wiener, N., autocorrelation function, 422.
- Wijsman, R., Bartlett decomposition, (Exercises 41.17-18) 263.
- Wilk, M. B., models in AV, 75; complete randomization, 81; expected MS in Latin squares, 138.
- Wilkinson, G. N., missing observations, 112-13.
- Wilks, S. S., LR test of independence of sets of variates (Exercises 41.10-11) 261-2, 271; homogeneity tests, (Example 42.2) 272, (Exercise 42.7) 282.
- Williams, E. J., discrimination, 326.
- Williams, W. H., unbiased regression-type estimators, 219, 223.
- Wilson, K. B., evolutionary operation, 158.
- Winters, P. R., exponential weighting, 501.
- Wise, J., autoregressive series, 417, (Exercise 47.17) 429.
- Wishart, J., distribution of multinormal covariances, 241-6, (Exercises 41.6-7) 261; non-central, 259; correlation between normal covariances, (Exercise 41.4) 260; distribution of covariance, (Exercise 41.13) 262; form of distribution of residual dispersion matrix, 275-6.
- Wold, H., moving average series, 415, 484; autoregressive series, 418, (Exercise 47.6) 427; causal models, 499.
- Wolfowitz, J., regression designs, 161; phases test, 354; rank serial correlation test, 360.
- Working, H., grouping in Markoff series, (Exercise 47.15) 428.
- Yates, F., method of weighted squares of means for two-way classification, 301, (Exercises 35.5-7) 53; missing observations, 111, 113; BIB designs, 142; inter-block information, 148; lattice designs, 153; confounding, 157; surveys, 166; sampling with unequal probabilities, 173; systematic sampling, 188; estimation of variance in multi-stage sampling, 199, 201, 204; efficiency of

- multi-stage sampling, 204; two-phase sampling, 228; domains of study, 229; sampling on successive occasions, (Exercises 40.9-10) 256-7.
- Youden, W. J., random balance experiments, 130; squares, 151-2.
- Young, D. H., quota sampling, (Exercise 40.11) 238.
- Yule, G. U., Slutsky-Yule effect, 378; equations for autoregressive series, 416; *see* Yule series.
- Yule series, correlogram and spectrum, (Example 47.8) 419-20; limiting case, (Example 47.10) 420-1; continuous analogue, (Example 47.11) 423; partial autocorrelations, 425; variance, (Exercise 47.8) 427; autocorrelations of residuals, (Exercise 47.10) 427; standard error of serial correlation, (Exercises 48.2, 48.14) 450, 452; serial correlations with errors of observation, (Exercise 49.11) 471; test of fit (Example 50.3) 480-1; multivariate, 496.
- Zaycoff, R., variate-difference method, 390.

DEPARTMENT OF STATISTICS
St Xavier's College
Calcutta.

